Evolution of Human Brain Size-Associated NOTCH2NL Genes Proceeds toward Reduced Protein Levels

Gerrald A. Lodewijk,[1] Diana P. Fernandes,[1] Iraklis Vretzakis,[1] Jeanne E. Savage,[2,3] and Frank M.J. Jacobs *[,1,3]

[1]Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands

[2]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, VU University, Amsterdam, The Netherlands

[3]Amsterdam Neuroscience, Complex Trait Genetics

***Corresponding author:** E-mail: f.m.j.jacobs@uva.nl.

Associate editor: Harmit Malik

Abstract

Ever since the availability of genomes from Neanderthals, Denisovans, and ancient humans, the field of evolutionary genomics has been searching for protein-coding variants that may hold clues to how our species evolved over the last ~600,000 years. In this study, we identify such variants in the human-specific *NOTCH2NL* gene family, which were recently identified as possible contributors to the evolutionary expansion of the human brain. We find evidence for the existence of unique protein-coding NOTCH2NL variants in Neanderthals and Denisovans which could affect their ability to activate Notch signaling. Furthermore, in the Neanderthal and Denisovan genomes, we find unusual *NOTCH2NL* configurations, not found in any of the modern human genomes analyzed. Finally, genetic analysis of archaic and modern humans reveals ongoing adaptive evolution of modern human *NOTCH2NL* genes, identifying three structural variants acting complementary to drive our genome to produce a lower dosage of NOTCH2NL protein. Because copy-number variations of the *1q21.1* locus, encompassing *NOTCH2NL genes*, are associated with severe neurological disorders, this seemingly contradicting drive toward low levels of NOTCH2NL protein indicates that the optimal dosage of NOTCH2NL may have not yet been settled in the human population.

Key words: **archaic genomes, brain size, human evolutionary genomics, human-specific genes, segmental duplications, Neanderthal, gene conversion.**

Introduction

The human brain tripled in size after we split from the common ancestor with our closest living relative species, the chimpanzees (Marino 1998; Herculano-Houzel 2009; Hofman 2014). The emergence of human-specific *NOTCH2NL* genes (Fiddes et al. 2018; Florio et al. 2018; Suzuki et al. 2018) coincided with this evolutionary expansion (Holloway et al. 2004; Pollen et al. 2015; Ju et al. 2016; Liu et al. 2017; Johnson et al. 2018; Kalebic et al. 2018) and their association to human brain development put *NOTCH2NL* genes forward as possible contributors to human's increased brain size. By enhancing Notch signaling, *NOTCH2NL* genes prolong proliferation of neuronal progenitor cells and expand cortical neurogenesis (Fiddes et al. 2018; Florio et al. 2018; Suzuki et al. 2018). *NOTCH2NL* genes are human specific and they emerged after a series of segmental duplications and gene conversion events involving the important neurodevelopmental gene *NOTCH2*. Four *NOTCH2NL* paralogs are present in modern humans: *NOTCH2NLA*, *NOTCH2NLB*, and *NOTCH2NLC* in the *1q21.1* locus (fig. 1A) and the pseudogene *NOTCH2NLR* next to the parental *NOTCH2* gene in the *1p12* locus. *NOTCH2NLB* represents the largest duplicon in the cluster, suggesting this was the first *NOTCH2NL* gene present in the genome (fig. 1B). Whereas copy-number variation is observed for *NOTCH2NLC* and *NOTCH2NLR* in the healthy human population; the copy number of *NOTCH2NLA* and *NOTCH2NLB* loci is highly stable in modern humans. In fact, *1q21.1* copy-number variations, mediated by breakpoints within the *NOTCH2NLA* and *NOTCH2NLB* genes, are associated with various neurological disorders (Brunetti-Pierri et al. 2008; Mefford et al. 2008; Bernier et al. 2016; Fiddes et al. 2018). These observations suggest that the total number of functional *NOTCH2NLA* and *NOTCH2NLB* alleles may be important for normal neuronal development. Given the highly variable genomic organization of the *1q21.1* locus, important questions remain about the level of variation in *NOTCH2NL* genes in the human population. In addition, it remains elusive whether the number and composition of *NOTCH2NL* genes has changed during recent human evolution. Here, we analyzed the segregation of coding variants in *NOTCH2NL* genes throughout human evolution and compared the composition of each *NOTCH2NL* locus between modern humans and archaic genomes. Our analysis revealed lineage-specific coding variants in each of the genomes of Neanderthals, Denisovans, and modern humans. Intriguingly, we find evidence for ongoing adaptive evolution of multiple structural variants in

Article

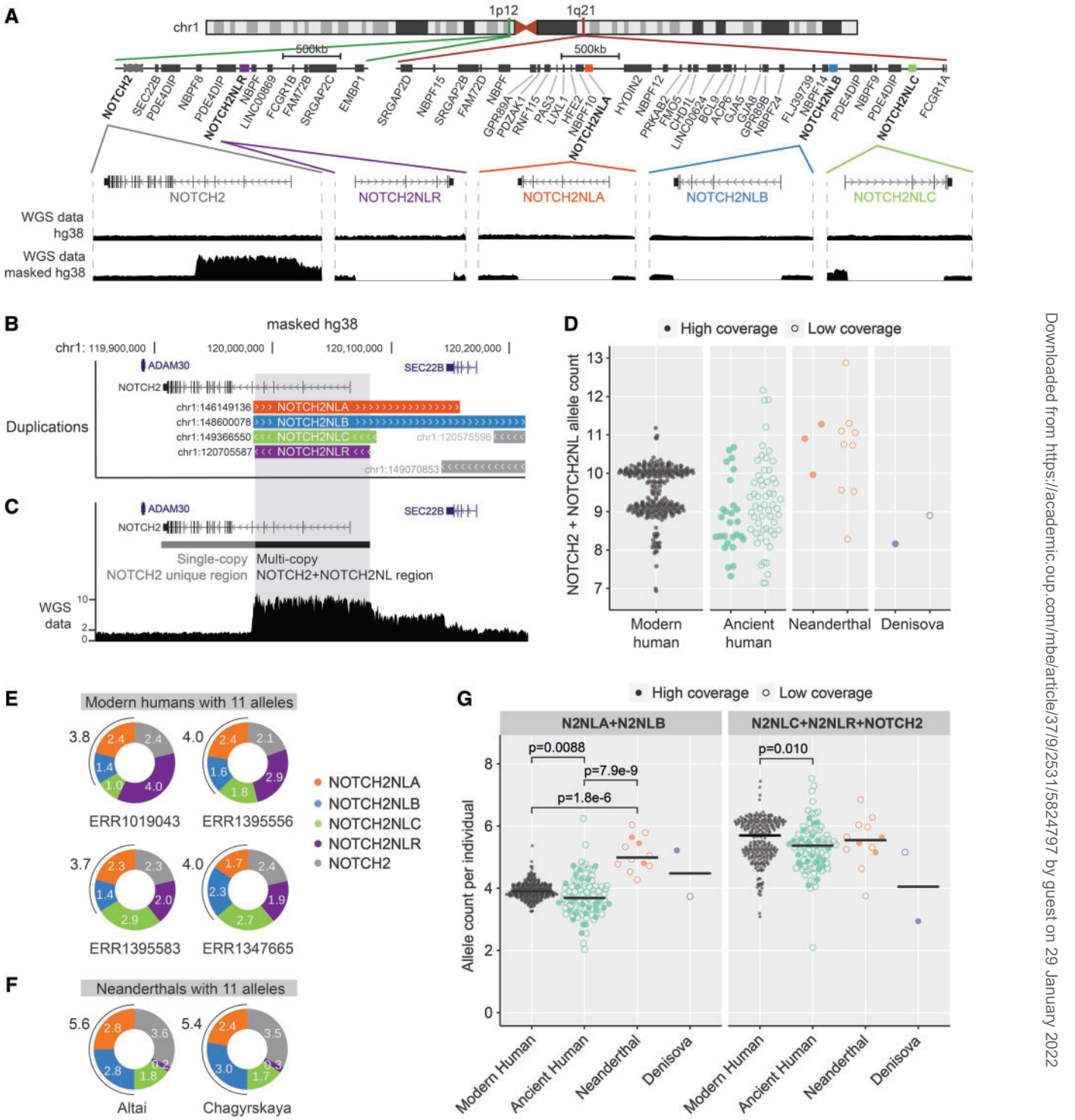

FIG. 1. *NOTCH2NL* copy-number analysis in modern human and archaic DNA samples. (*A*) Overview of *NOTCH2* and *NOTCH2NL* loci in the human genome (hg38). Zoom-ins show sequence read depth at the different loci of data mapped on hg38 or masked hg38 reference genome. (*B*) Tracks showing *NOTCH2NL* duplicons from the segmental UCSC browser duplication track in the *NOTCH2* locus. (*C*) Example showing *NOTCH2*- and *NOTCH2NL*-derived sequencing reads piled up on the *NOTCH2* locus on the masked hg38 genome. (*D*) Quantification of *NOTCH2* + *NOTCH2NL* alleles per individual using relative coverage of multicopy/single-copy regions. Modern human, $n = 279$. Ancient human: high ($n = 27$)/low ($n = 53$) coverage; Neanderthal high ($n = 3$)/low ($n = 9$) coverage; Denisova high ($n = 1$)/low ($n = 1$) coverage. (*E*, *F*) *NOTCH2NL* allele counts estimated from the average density of paralog-specific SUNs in modern human outliers (*E*) and Neanderthals (*F*) showing evidence for the presence of 11 alleles in total (two alleles *NOTCH2* + nine alleles *NOTCH2NL*). (*G*) Comparison of allele count grouped by *NOTCH2NLA* + *NOTCH2NLB* (Kruskal–Wallis $P = 1.8e\text{-}8$), and *NOTCH2NLR* + *NOTCH2NLC* + *NOTCH2* (Kruskal–Wallis $P = 0.0055$). Kruskal–Wallis test was followed up by Dunn's test, significant comparisons are indicated in the plots. Modern human, $N = 279$; ancient human, $N = 80$; Neanderthal, $N = 12$; and Denisova, $N = 2$.

modern human *NOTCH2NL* genes, acting in synergy and complementary to drive our genome to produce a lower dosage of NOTCH2NL protein. The evolutionary forces mediated by gene conversion Chen et al. 2007, which we find is still ongoing between *NOTCH2NL* loci at a high frequency in modern humans, exemplify how recently duplicated regions in our genome can undergo rapid structural evolution to reach an optimal configuration and functionality. For humans, this may have had important consequences for how a key developmental process such as Notch signaling has evolved in the period after the emergence of *NOTCH2NL* genes and the changes they effectuated on human brain development.

Additional Copies of *NOTCH2NLA* or *NOTCH2NLB* in Neanderthals

To assess the structural evolution of each of the *NOTCH2NL* loci throughout human evolution, we first assessed the structural variability of *NOTCH2NL* loci in the modern human population. Previous estimations of total *NOTCH2NL* copy number in individuals could not efficiently distinguish between paralogous *NOTCH2NL* loci subject to recent ectopic gene conversion, as observed between *NOTCH2-NOTCH2NLR* and between *NOTCH2NLA-NOTCH2NLB* (Dougherty et al. 2017; Fiddes et al. 2018). Here, we used an alternative strategy that takes into account gene conversion between paralogous *NOTCH2NL* loci: For each genome, we assessed total number of *NOTCH2NL* alleles based on sequence read coverage and matched this with information about the presence or absence of *NOTCH2NL*-paralog identifying single-unique nucleotides (SUNs) (Sudmant et al. 2010). This provides an accurate assessment of the absolute number of *NOTCH2NL* alleles in each individual genome and a detailed overview of the structural variability of *NOTCH2NL* genes as a consequence of gene conversion (supplementary tables S1–S5, Supplementary Material online). We verified the accuracy of our methodology by showing concordance with previous *NOTCH2NL* assembly–based estimations (supplementary table S6, Supplementary Material online). To assess the total number of *NOTCH2NL* alleles across the human population, the genomes of 279 individuals from the Simons diversity data set (Mallick et al. 2016) were mapped onto a modified hg38 genome in which the *NOTCH2NL* loci are masked (fig. 1*A*). On this modified hg38 genome, all *NOTCH2NL*-derived reads map onto the 5′ side of the *NOTCH2* locus, the part of *NOTCH2* that was originally duplicated forming the *NOTCH2NL* genes (fig. 1*B* and *C*). The coverage analysis reveals that the majority of the human population has ten alleles, encompassing two alleles from *NOTCH2* and two alleles from each of the four *NOTCH2NL* loci (fig. 1*D*). Using the combined sequence coverage and SUN analysis, we determined that each individual contained 4 alleles combined of the highly similar *NOTCH2NLA* and *NOTCH2NLB* genes. The individuals that have nine, eight, or seven alleles were all confirmed as hetero- or homozygotic for *NOTCH2NLC* and *NOTCH2NLR* (supplementary fig. S1*A* and *B*, Supplementary Material online). Four human individuals have one extra allele of *NOTCH2NLC* or *NOTCH2NLR*, indicating that *NOTCH2NL* duplications happen in the healthy human population (fig. 1*E*). Next, we analyzed genomes of ancient humans (0.1k–45k years old) (Keller et al. 2012; Fu et al. 2014, 2016; Gamba et al. 2014; Lazaridis et al. 2014; Olalde et al. 2014; Raghavan et al. 2014; Rasmussen et al. 2014, 2015; Seguin-Orlando et al. 2014; Skoglund et al. 2014, 2017; Günther et al. 2015, 2018; Jones et al. 2015, 2017; Cassidy et al. 2016; Martiniano et al. 2016; Schiffels et al. 2016; Saag et al. 2017; Bhattacharya et al. 2018; de la Fuente et al. 2018; Krzewińska et al. 2018; Valdiosera et al. 2018; Wright et al. 2018; Sánchez-Quinto et al. 2019), Neanderthals (38k–100k years old) (Green et al. 2010; Prüfer et al. 2014, 2017; Hajdinjak et al. 2018; Slon et al. 2018; Mafessoni et al. 2020), and Denisovans (64k–100k years old) (Meyer et al. 2012; Slon et al. 2017). Although most of the ancient human genomes display *NOTCH2NL* allele numbers that fall within the range of modern humans, several of the 12 available Neanderthal genomes show increased coverage, which indicates they contained an extra *NOTCH2NL* duplication (fig. 1*D*). Whereas the combined copy number of *NOTCH2NLA* and *NOTCH2NLB* is highly stable in healthy modern humans, SUN-based copy-number estimation suggests that Neanderthals carried an extra duplication of the *NOTCH2NLA* or *NOTCH2NLB* gene (fig. 1*F* and *G* and supplementary fig. S1*C*, Supplementary Material online). Whether this is a gain in Neanderthal or a loss in modern humans remains elusive. In addition, all Neanderthal genomes showed evidence of extensive gene conversion between *NOTCH2* and *NOTCH2NLR* (supplementary fig. S1*C*, Supplementary Material online), a phenomenon observed only occasionally in modern humans (supplementary fig. S1*D* and *E*, Supplementary Material online).

Neanderthals and Denisovans Carried Specific NOTCH2NL Variants

We next investigated whether the archaic genomes contained any coding sequence variants that may have encoded unique NOTCH2NL protein variants. Despite an overall high similarity (99.9%) between human and Neanderthal/Denisovan *NOTCH2NL* exons, we found evidence for two Neanderthal-specific coding variants and one Denisova-specific coding variant (fig. 2*A*). In the Altai Neanderthal genome, an ATG > ATA (M40I) missense variant ($NOTCH2NL^{Nea\text{-}M40I}$) is detected in 17/242 (∼8%) of the sequencing reads corresponding to one allele out of the nine *NOTCH2NL* alleles found in Altai Neanderthals. The second Neanderthal-specific variant is a N232S missense variant ($NOTCH2NL^{Nea\text{-}N232S}$) detected in 28/177 (∼18%) of sequencing reads, corresponding to two alleles. This variant is also present in the genomes of the Vindija and Chagyrskaya Neanderthals and most of the low-coverage Neanderthal genomes, indicating the $NOTCH2NL^{Nea\text{-}N232S}$ variant was a common variant in the Neanderthal lineage. In the Denisova3 genome, a Denisovan-specific E258A missense variant ($NOTCH2NL^{Den\text{-}E258A}$) is found in 38/203 (∼19%) of the sequencing reads, also corresponding to two alleles. Importantly, none of these variants are found in the 279 modern human genomes of the Simons diversity data set. Interestingly, the $NOTCH2NL^{Nea\text{-}N232S}$ was found as a rare

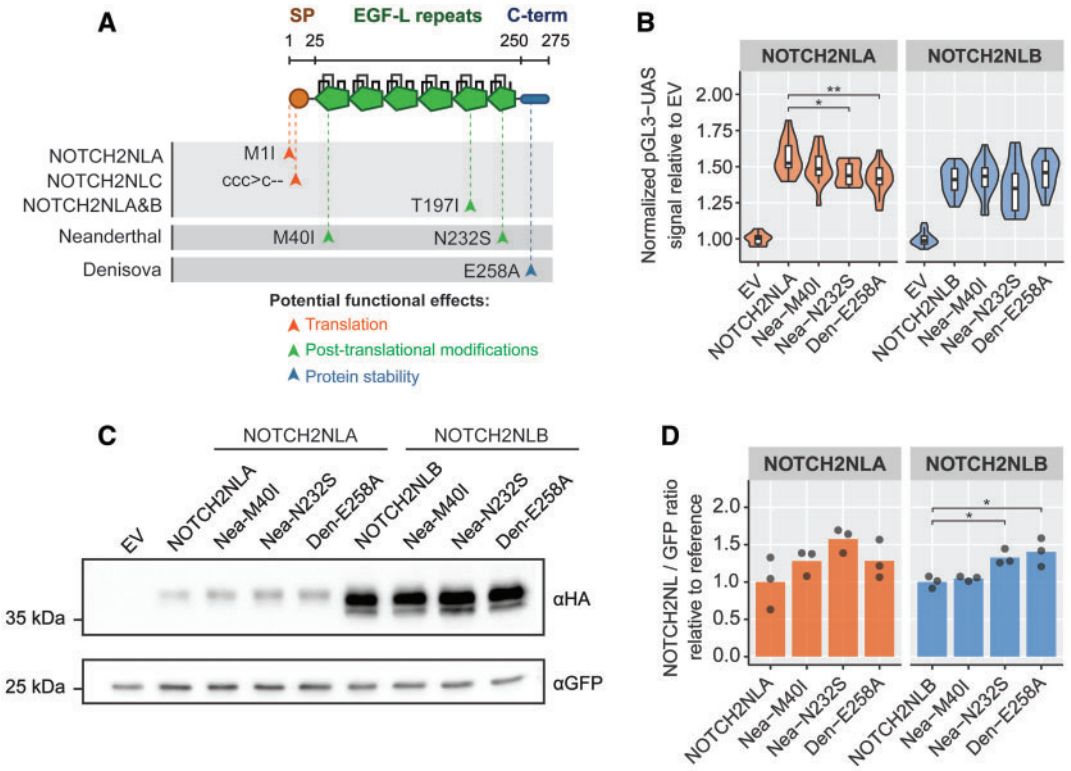

Fig. 2. Characterization of archaic *NOTCH2NL* coding variants. (*A*) Overview of modern human, Neanderthal-specific, and Denisovan-specific coding variants. (*B*) Coculture NOTCH2 reporter assay testing Neanderthal and Denisovan variants reconstructed in the human *NOTCH2NLA* cDNA ($n = 15$ in three experiments, analysis of variance (ANOVA) $P = 0.002$, followed by Tukey's test), or the human *NOTCH2NLB* cDNA ($n = 20$ in four experiments, ANOVA $P = 0.07$). (*C*) Western blot analysis of Neanderthal and Denisovan variants. Plasmids were transfected in equimolar amounts. (*D*) Quantification of protein level from three independent experiments for NOTCH2NLA (ANOVA $P = 0.12$) and NOTCH2NLB (ANOVA $P = 0.006$, followed by Tukey's test). Asterisks indicate significant values from Tukey's post hoc tests: $^*P < 0.05$ and $^{**}P < 0.01$.

variant in modern humans (rs375605753) with an allele frequency of 0.0002 in UK Biobank exome sequencing data ($N = 49{,}593$), suggesting this was one of the Neanderthal-derived genetic variants that was contributed to the human genome after interbreeding with Neanderthals (Dannemann and Racimo 2018). It should be noted that the highly fragmented assemblies of archaic genomes prevents us from making solid claims about which *NOTCH2NL* paralog each of these archaic variants reside in. Taking this into account, we assessed the potential functional implications of the Neanderthal and Denisova variants by reconstructing the archaic NOTCH2NL variants in NOTCH2NLA and NOTCH2NLB for functional testing in a previously established Notch signaling reporter assay (Groot et al. 2014; Habets et al. 2015; Fiddes et al. 2018) (supplementary fig. S2*A*, Supplementary Material online). Surprisingly, the introduction of the Nea-N232S and Den-E258A into human NOTCH2NLA showed a modest but significant decrease in potency to enhance Notch signaling (fig. 2*B*). To find an explanation for the functional divergence of the archaic NOTCH2NL variants, we investigated the potential structural implications in more detail (supplementary fig. S2*B*, Supplementary Material online). The Neanderthal M40I variant is located in EGF-L domain 1 and disrupts the predicted start codon of NOTCH2NLA. The Neanderthal N232S variant is located in EGF-L domain 6, which is fully conserved between NOTCH paralogs and between species (supplementary fig. S2*C*, Supplementary Material online). The N232 residue is part of an important motif for glycosylation, a posttranslational modification which mediates EGF-L folding (Takeuchi et al. 2017) and NOTCH–ligand interactions (Jafar-Nejad et al. 2010) (supplementary fig. S2*D*, Supplementary Material online). As such, the N232S variant is predicted to alter NOTCH2NL protein interaction dynamics or protein stability (supplementary fig. S2*E*, Supplementary Material online). Indeed, the corresponding rare single-nucleotide polymorphism (SNP) in modern humans (rs375605753) is predicted to be deleterious (Pejaver et al. 2017). The Denisova E258A variant is located in the C-terminal domain of NOTCH2NL, an intrinsically disordered region known to play a role in protein stability (Duan et al. 2003; Fiddes et al. 2018). Analysis using IUPred2A (Mészáros et al. 2018) suggests that this substitution alters the state of the NOTCH2NL C-terminal domain, potentially affecting protein stability (supplementary fig. S2*E* and *F*, Supplementary Material online). In support of this, a modest increase in protein level was observed for the Den-E258A and Nea-N232S variants introduced into human NOTCH2NLB (fig. 2*C* and *D*). This suggests that these archaic variants positively affected protein translation or stability. Altogether, Denisovans and Neanderthals carried

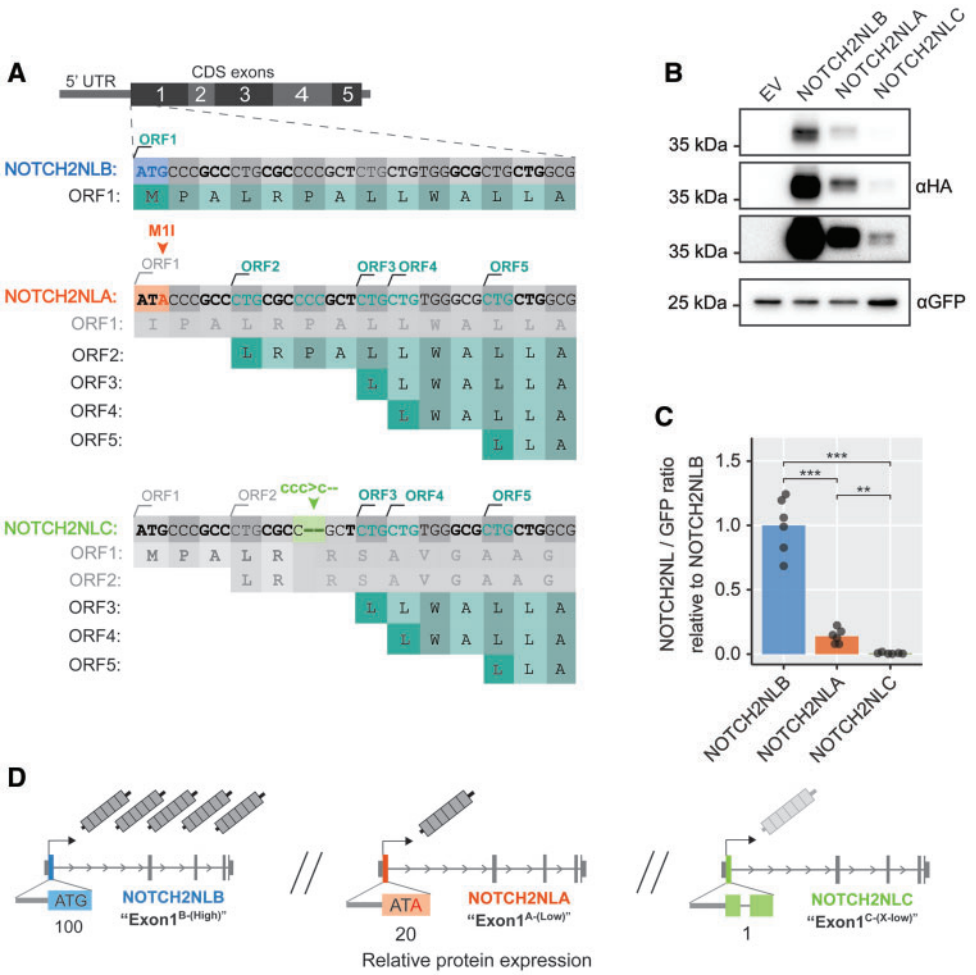

FIG. 3. *NOTCH2NL* Exon1 variants define protein expression level. (*A*) Overview of *NOTCH2NL* Exon1 variants in *NOTCH2NLB* (blue), *NOTCH2NLA* (orange), and *NOTCH2NLC* (green). The ORFs produced by each variant are indicated in dark green. (*B*) Western blot analysis of *NOTCH2NL* Exon1 coding variants. (*C*) Quantification of protein expression level from equimolar quantities of *NOTCH2NLB*, *NOTCH2NLA*, or *NOTCH2NLC* full-length cDNAs. Data from six independent experiments, analysis of variance (ANOVA) (Welch corrected) $P = 2.7e\text{-}05$, followed Games–Howell test: $^{**}P < 0.01$ and $^{***}P < 0.001$. (*D*) Overview of NOTCH2NL loci, the configuration of the Exon1 variants, and the relative levels of NOTCH2NL protein they produce.

alleles in their genome which are likely to have affected the function of their *NOTCH2NL* genes.

Variants in Exon1 of *NOTCH2NL* Genes Determine NOTCH2NL Protein Levels

Unexpectedly, we noticed that the NOTCH2NLA$^{Nea\text{-}M40I}$ variant, predicted to lack the first 83 amino acids, was not different in size from NOTCH2NLB. Likewise, no decrease in protein size was observed for NOTCH2NLA, predicted to lack the first 39 amino acids. Analysis of multiple 5′ truncated *NOTCH2NL* cDNAs reveals that instead of the conventional ATG initiation sites on positions M40 and M84, multiple unconventional CTG start sites in the 5′ side of *NOTCH2NL* drive translation of NOTCH2NLA and NOTCH2NLA$^{Nea\text{-}M40I}$ proteins (Kearse and Wilusz 2017) (fig. 3*A* and supplementary fig. S3*A*–*G*, Supplementary Material online). As a result and as opposed to what is predicted by gene models, human *NOTCH2NLA* and Neanderthal *NOTCH2NLA*$^{Nea\text{-}M40I}$ encode almost full-length NOTCH2NL proteins with a functionally intact N-terminal signal peptide. Importantly, our analysis also reveals that the usage of unconventional translation initiation sites has major consequences for the level of NOTCH2NL protein produced by each of the *NOTCH2NL* genes. *NOTCH2NLA*, which lacks the first start codon produces a 5-fold lower level of NOTCH2NL protein compared with NOTCH2NLB (fig. 3*A*–*C*). *NOTCH2NLC* is also forced to use downstream CTG sites for translation initiation and gives rise to normal-sized NOTCH2NL protein (fig. 3*B*). However, due to the combination of the *NOTCH2NLC*-characteristic 2-bp deletion and upstream open-reading frames (ORFs), the expression level of NOTCH2NLC is extremely low, at only 1% compared with NOTCH2NLB (fig. 3*C*). These new insights reveal that the level of NOTCH2NL protein generated by each of the genes is predominantly dependent on the presence or absence of three specific coding variants in Exon1 (fig. 3*D*). Compared with the *NOTCH2NLB* configuration of Exon1 (Exon1$^{B\text{-}(High)}$-variant) which produces high levels of NOTCH2NL protein, the M1I substitution in NOTCH2NLA (Exon1$^{A\text{-}(Low)}$-variant) produces 5-fold less NOTCH2NL protein. The configuration of *NOTCH2NLC*, which has the 2-bp

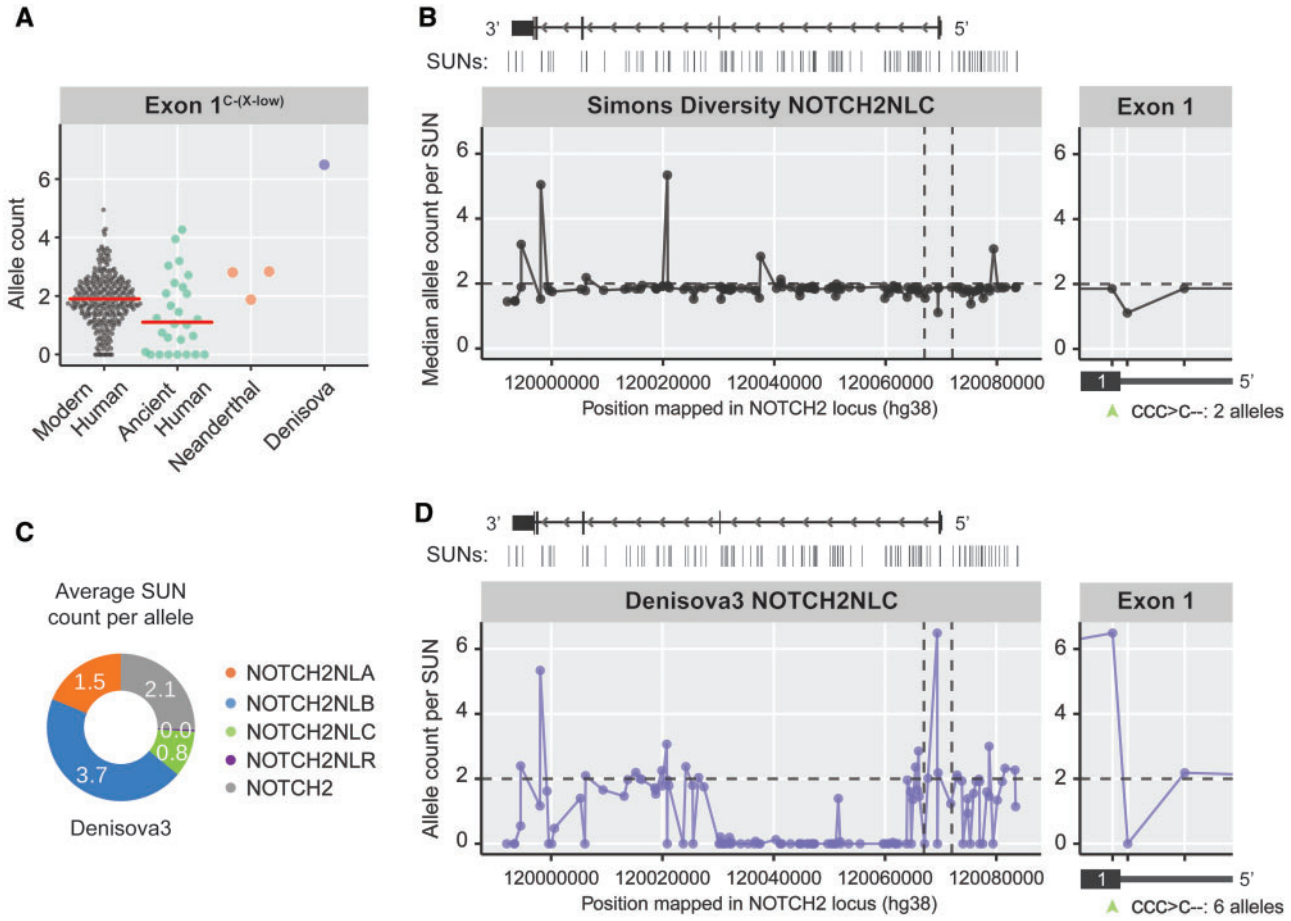

FIG. 4. *NOTCH2NLC* configuration in Denisova3 compared with modern humans. (*A*) Plot showing the Exon1$^{C-(X-low)}$ allele count for modern humans, ancient humans, Neanderthals, and Denisovan. Note the unusual allele count for Denisovan. (*B*) Modern human's median allele count plotted for each of the *NOTCH2NLC*-specific SUNs distributed along the *NOTCH2NL* locus. Vertical dashed lines indicate the region around Exon1. Zoom-in shows SUN count in Exon1, including the Exon1$^{C-(X-low)}$ variant indicated by green arrowhead. (*C*) *NOTCH2NL* allele counts in the Denisova3 genome, estimated from the average density of paralog-specific SUNs. (*D*) Denisova3 allele count plotted for each of the *NOTCH2NLC*-specific SUNs distributed along the *NOTCH2NL* locus. Zoom-in shows *NOTCH2NLC* SUN count in Exon1, including the Exon1$^{C-(X-low)}$ variant as indicated by green arrowhead.

deletion in Exon1, (Exon1$^{C-(X-low)}$-variant) results in extremely low levels of NOTCH2NL protein. Importantly, ectopic gene conversion between *NOTCH2NL* loci can result in transfer of Exon1-variants from one *NOTCH2NL* gene to another. As a consequence, the total dosage of NOTCH2NL protein in each individual may not be defined by the copy number of each of the *NOTCH2NL* genes, but by the level of Exon1-variant carry-over via gene conversion between *NOTCH2NL* genes.

Unusual Configuration of *NOTCH2NL* Genes in the Denisova3 Genome

To assess the extent to which gene conversion influences the distribution of Exon1-variants between *NOTCH2NL* genes, we investigated the distribution of SUNs across the *NOTCH2NL* loci. First, we analyzed modern human *NOTCH2NLC* for evidence of gene conversion. Analysis of the Exon1 configuration of *NOTCH2NL* genes reveals that most modern humans contain two *NOTCH2NLC*-derived Exon1$^{C-(X-low)}$-variants (fig. 4*A*), present in both alleles of *NOTCH2NLC*. Furthermore, an equal distribution was found for *NOTCH2NLC* SUNs across the *NOTCH2NL* locus in most modern human individuals (fig. 4*B*), suggesting that gene conversion between *NOTCH2NLC* and other *NOTCH2NL* loci does not commonly happen. A similar pattern was found in Neanderthals and ancient humans (fig. 4*A* and supplementary fig. S4*A* and *B*, Supplementary Material online). This indicates that the majority of Neanderthal, archaic human, and modern human genomes have two *NOTCH2NLC* alleles carrying the Exon1$^{C-(X-low)}$-variant. The Denisova3 genome however, shows a strikingly different pattern.

The presence of *NOTCH2NL*-paralog-specific SUNs across the *NOTCH2NL* loci shows that *NOTCH2NLA*, *NOTCH2NLB*, and *NOTCH2NLC* genes are present in the Denisova3 genome (fig. 4*C*). Based on the complete absence of *NOTCH2NLR* SUNs and a total coverage representative of only six *NOTCH2NL* alleles (fig. 1*D*), it is likely the Denisova3 genome had a homozygous deletion of *NOTCH2NLR*. Remarkably, despite good coverage of the Exon1 region in the Denisova3 genome (36X), all *NOTCH2NL*-derived reads from Exon1 carry the *NOTCH2NLC*-derived Exon1$^{C-(X-low)}$-variant (fig. 4*D*). This implies that all six Denisovan *NOTCH2NL* alleles produced NOTCH2NL protein at an extremely low level.

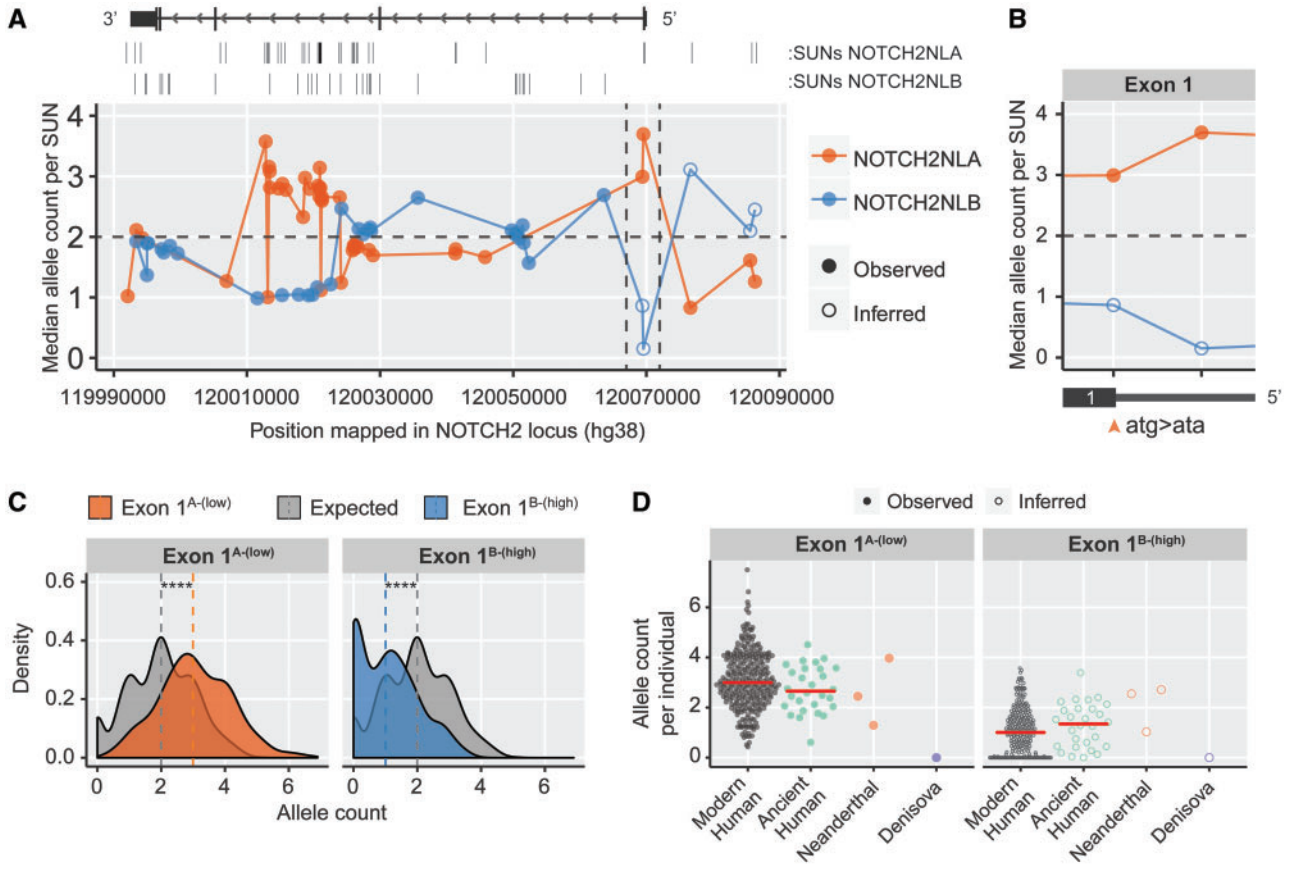

FIG. 5. Exon1 variant frequencies in modern human and ancient genomes. (*A*) Median allele count for each of the *NOTCH2NLA*- and *NOTCH2NLB*-specific SUNs along the *NOTCH2NL* locus in Simons diversity genomes ($N = 279$). (*B*) Zoomed in region of Exon1, orange arrowhead indicates Exon1$^{B\text{-}(High)}$ (ATG)/Exon1$^{A\text{-}(Low)}$ (ATA) variant positions. (*C*) Distribution of Exon1$^{A\text{-}(Low)}$ and Exon1$^{B\text{-}(High)}$ (inferred) variants in Simons diversity genomes. Expected distribution models equal frequency of both variants. Vertical dashed lines indicate medians. $N = 279$, Kolmogorov–Smirnov test: $P < 2e\text{-}16$. (*D*) Analysis of Exon1$^{A\text{-}(Low)}$ and Exon1$^{B\text{-}(High)}$ (inferred) variant frequency in modern humans and archaic genomes. Red lines indicate medians.

Unfortunately, the lack of other high-coverage Denisovan genomes prevents us from assessing whether this is an individual-specific genotype or whether similar *NOTCH2NLC* gene conversions were frequent in the Denisovan population. Importantly, this pattern of Exon1$^{C\text{-}(X\text{-}low)}$-variant distribution in Denisovan *NOTCH2NL* genes, or anything similar to it, has not been observed in any of the analyzed genomes of Neanderthals or healthy modern humans (supplementary fig. S4*C*, Supplementary Material online).

Evolution of Modern Human *NOTCH2NL* Genes Trends toward Lower NOTCH2NL Levels

Even though *NOTCH2NLA* and *NOTCH2NLB* are capable of producing a structurally similar NOTCH2NL protein, the protein levels they produce differ by 5-fold. In the SUN analysis, we find evidence of extensive gene conversion between the *NOTCH2NLA* and *NOTCH2NLB* loci: The median SUN depth shifts in favor of either allele in different regions of the loci, indicating that parts of the *NOTCH2NLA*-sequence are frequently overwritten by *NOTCH2NLB*-sequence and vice versa (fig. 5*A*). Most regions with a strong shift in distribution of *NOTCH2NLA* or *NOTCH2NLB* SUNs are intronic, not predicted to impact the structure and level of NOTCH2NL protein. However, the configuration of Exon1 in *NOTCH2NLA* and *NOTCH2NLB* shows a median allele depth strongly in favor of the Exon1$^{A\text{-}(Low)}$-variant (fig. 5*B*). This is striking because it suggests that the vast majority of the population carries three or four alleles with the *NOTCH2NLA*-derived Exon1$^{A\text{-}(Low)}$-variant and only one or zero alleles with the *NOTCH2NLB*-derived Exon1$^{B\text{-}(High)}$-variant (fig. 5*C*). The shift in Exon1$^{A\text{-}(Low)}$-variant distribution was confirmed in 49,593 exomes from the UK Biobank (Van Hout et al. 2019) (supplementary fig. S5*A*, Supplementary Material online) and was also observed in the genomes of ancient modern humans (fig. 5*D*). The observed imbalance in distribution of Exon1-variants indicates that the Exon1$^{B\text{-}(High)}$-variant, producing the highest levels of NOTCH2NL protein, is being lost or actively being purged out from the modern human population by gene conversion. The increase of the Exon1A-(Low) variant frequency to three or four alleles per individual is likely caused by gene conversion between the *NOTCH2NLA* and *NOTCH2NLB* loci, which can occur during meiosis or in early embryonic development for very unstable loci (Chen et al. 2007; Bruder et al. 2008; Vadgama et al. 2019).

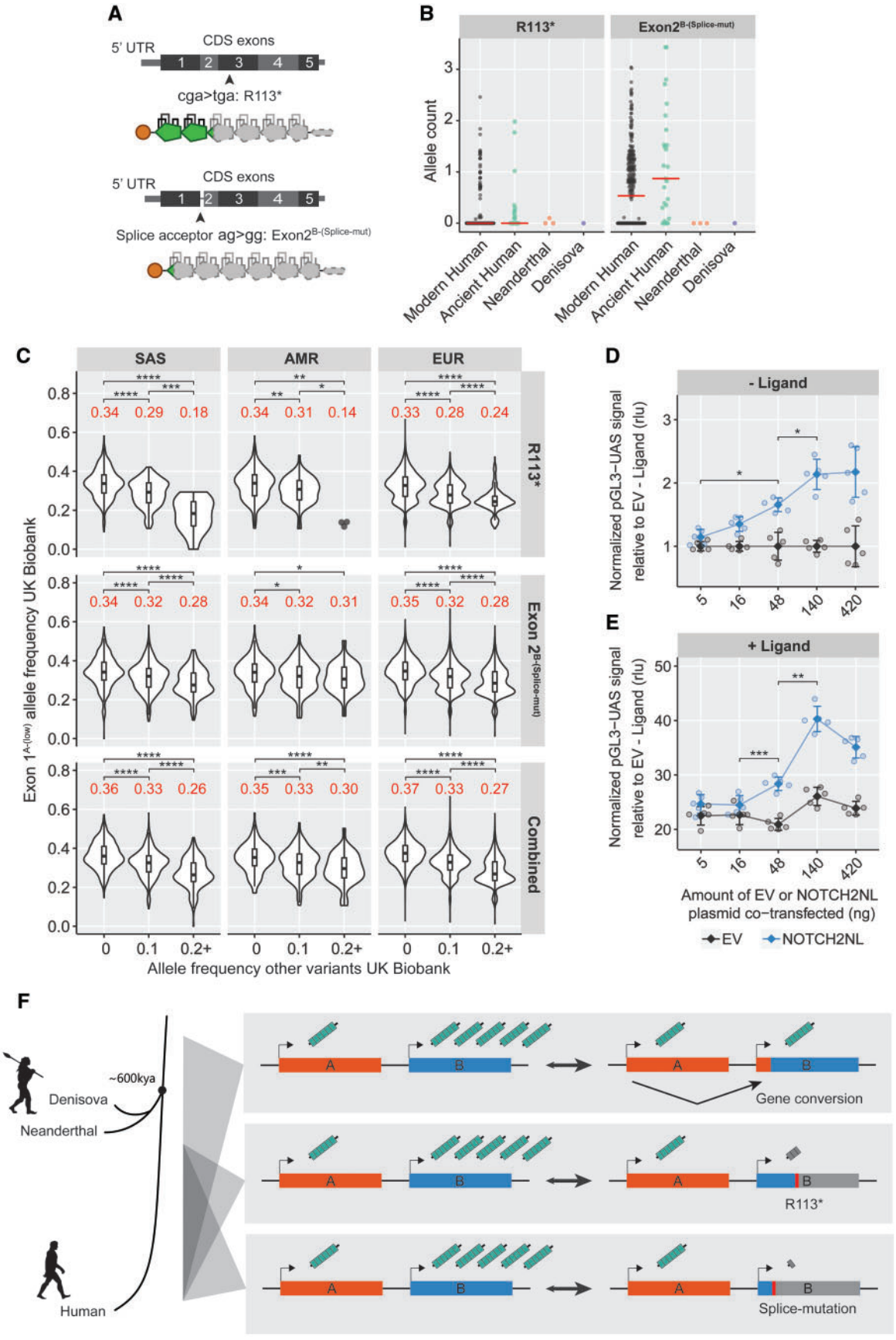

FIG. 6. Additional deleterious *NOTCH2NL* variants are present specifically in humans. (*A*) Overview of the R113* and Exon$^{2B\text{-}(Splice\text{-}mut)}$ deleterious variants on NOTCH2NL protein structure. (*B*) R113* and Exon$^{2B\text{-}(Splice\text{-}mut)}$ allele count in modern human and archaic genomes. (*C*) UK Biobank data for SAS, AMR, and EUR ancestries showing association of Exon1$^{A\text{-}(Low)}$ frequency with R113* frequency, Exon2$^{B\text{-}(Splice\text{-}mut)}$ frequency, and their

Spreading of Modern Human-Specific Deleterious Variants Indicates Strong Compensatory Mechanisms

Despite the relatively high abundance of Exon1$^{A\text{-}(Low)}$ variants in *NOTCH2NLA* and *NOTCH2NLB*, some individuals still carry a relatively high number of Exon1$^{B\text{-}(High)}$ variants. We found that individuals with a relatively high number of the Exon1$^{B\text{-}(High)}$ variant and low number of the Exon1$^{A\text{-}(Low)}$ variant often carry a nonsense SNP (R113*) in NOTCH2NLB, which leads to a premature stop-codon and a severely truncated NOTCH2NL protein (fig. 6*A*). In addition, we found another variant in the splice acceptor sequence of exon 2 (Exon$^{2B\text{-}(Splice\text{-}mut)}$) (fig. 6*A* and supplementary fig. S6*A*, Supplementary Material online). This variant falls outside the coding region and therefore was not detected before. The AG > GG mutation is predicted to lead to an alternative splicing event, resulting in a frameshift and truncation of NOTCH2NL proteins at amino acid 30 (Dougherty et al. 2018). On hg38, this variant is annotated in *NOTCH2NLB* and it is present at a high allelic frequency in human genomes from the Simons diversity data (supplementary fig. S6*B*, Supplementary Material online) and the UK Biobank (supplementary fig. S6*C*, Supplementary Material online). The R113* variant is less frequently observed. Surprisingly, the splice acceptor variant Exon$^{2B\text{-}(Splice\text{-}mut)}$ and the R113* mutation were not found in any of the currently available Neanderthal or Denisovan genomes (fig. 6*B* and supplementary fig. S6*B*, Supplementary Material online) and are therefore recently evolved human lineage-specific adaptations. Both loss-of-function variants appear to be common in the South-Asia (SAS), American (AMR), and European (EUR) ancestries and are only sporadically present in East-Asian (EAS) or African (AFR) ancestries in the UK Biobank data (fig. 6*B* and supplementary fig S6*C* and *D*, Supplementary Material online). Segregation of the disruptive alleles appeared to be nonrandom because we found a clear correlation between the individual's number of Exon1 $^{A\text{-}(low)}$ or Exon1$^{B\text{-}(High)}$ variants and the presence of disruptive R113* and Exon$^{2B\text{-}(Splice\text{-}mut)}$ mutations: Individuals with a relatively high number of the Exon1$^{B\text{-}(High)}$ variant, often carry one or two alleles of the disruptive R113* mutation in *NOTCH2NLB* (fig. 6*C*-upper panel and supplementary fig. S6*C* and *D*, Supplementary Material online). A strikingly similar pattern was observed for the Exon$^{2B\text{-}(Splice\text{-}mut)}$ mutation (fig. 6*C*-middle panel and supplementary fig. S6*C* and *D*, Supplementary Material online). Conversely, individuals with a relatively higher number of Exon1$^{A\text{-}(Low)}$ variants are more likely to lack either the R113* or splice acceptor mutations in *NOTCH2NLB* (fig. 6*C*-lower panel and supplementary fig. S6*C* and *D*, Supplementary Material online). In the EAS population, the more sporadic occurrence of both disruptive NOTCH2NL variants correlates with an overall higher Exon1$^{A\text{-}(Low)}$ frequency instead (supplementary fig. S6*D* and *E*, Supplementary Material online). This reveals a complex pattern of *NOTCH2NL* configurations, where multiple structural variants in *NOTCH2NLB*, the gene that has the largest contribution to the overall NOTCH2NL levels, seem to act complementary to reduce NOTCH2NL protein levels. In the Simons diversity data set, we observe highly similar patterns, but this analysis lacked statistical power due to the relatively small sample size per ancestry group (supplementary fig. S7*A*–*C*, Supplementary Material online). Taken together, our findings suggest that a relatively high load of the Exon1$^{B\text{-}(High)}$ variant often co-occurs with the presence of nonsense variants in *NOTCH2NLB*. Our data suggest that on the individual's genome level, gene conversion of the Exon1$^{B\text{-}(High)}$ variant into the Exon1$^{A\text{-}(Low)}$ variant acts in concert with nonsense variants in *NOTCH2NLB* to reduce overall NOTCH2NL protein level. This seems particularly relevant because we observe a strong dosage-dependent effect of NOTCH2NL on Notch signaling activation (fig. 6*D* and *E*), indicating that NOTCH2NL dosage is tightly associated with its functional output, which in the brain is controlling cortical neurogenesis. Altogether, the identification of Neanderthal-, Denisovan-, and modern human-specific coding variants and their complementary functional impact on NOTCH2NL protein levels suggests that the optimal level of NOTCH2NL protein has been under strong selective pressure in recent human evolution and is still being optimized in the human population (fig. 6*F*).

Discussion

The detection of multiple lineage-specific coding variants and the rapid spread of some of them throughout modern human genomes shows that the structure of human *NOTCH2NL* genes has been subject to ongoing adaptive evolution since the split of modern humans, Neanderthals, and Denisovans from our common ancestor ~600,000 years ago. This is corroborated by the presence of additional copies of *NOTCH2NLA* or *NOTCH2NLB* in Neanderthal genomes and the unusual configuration of six *NOTCH2NLC*-derived Exon1$^{C\text{-}(X\text{-}low)}$ variants in the Denisova3 genome. Notably, none of the 279 modern human individuals analyzed in detail in this study showed similar configurations and it is questionable whether such configurations are found in the healthy human population. This raises questions about the health state of the juvenile Denisovan female from the Denisova3

Fig.6. Continued

combined total grouped by ancestry. R113* Kruskal–Wallis: SAS $P = 2.2e\text{-}16$, AMR $P = 7.8e\text{-}5$, and EUR $P = 2.2e\text{-}16$. Exon2$^{B\text{-}(Splice\text{-}mut)}$ Kruskal–Wallis: SAS $P = 4.6e\text{-}15$, AMR $P = 0.04$, and EUR $P = 1.1e\text{-}15$. Combined Kruskal–Wallis: SAS $P = 2.2e\text{-}16$, AMR $P = 9.9e\text{-}7$, and EUR $P = 2.2e\text{-}16$. Significant groups were followed by Dunn's test. (*D*, *E*) Dose–response curve using increasing amounts of NOTCH2NL in the coculture NOTCH2 reporter assay. (*D*) NOTCH2 expressing cells are cocultured with U2OS (− ligand, analysis of variance [ANOVA] $P = 7.4e\text{-}7$, followed by Tukey's test: $^*P < 0.05$) or (*E*) U2OS-JAG2 (+ ligand, ANOVA $P = 4.7e\text{-}9$, followed by Tukey's test: $^{**}P < 0.01$ and $^{***}P < 0.001$) cells. $n = 5$ per condition, displayed as mean ± SD. (*F*) General overview schematic showing the impact of variants in NOTCH2NL genes on the production of NOTCH2NL protein and the time/lineage where they were segregating. Asterisks indicate significant values from Dunn's post hoc tests: $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, and $^{****}P < 0.0001$. EAS $N = 266$, SAS $N = 1{,}174$, AMR $N = 444$, EUR $N = 46{,}578$, and AFR $N = 1{,}087$.

genome, but because the DNA was isolated from a finger bone, information about her physical condition or cause of death is lacking (Meyer et al. 2012). Although our data indicate no major role for *NOTCH2NLC* in normal development due to its low protein expression levels and common loss of one allele, recent studies describe repeat expansions in the 5′UTR *NOTCH2NLC* genes linked to neurodegenerative disorders (Deng et al. 2019; Ishiura et al. 2019; Okubo et al. 2019; Sone et al. 2019; Hayashi et al. 2020; Jiao et al. 2020; Sun et al. 2020). So, it is possible that this repeat expansion leads to disease via a gain-of-function mechanism. For example, it could be that the repeat expansions in *NOTCH2NLC* lead to an N-terminally extended ORF, which in turn may cause aberrantly high expression of NOTCH2NL or production of toxic NOTCH2NL protein variants. Further experiments regarding these possibilities are necessary to understand the mechanisms that underlie the reported disease phenotypes.

Our data suggest that gene conversion still plays a central role in exchanging coding variants between *NOTCH2NLA* and *NOTCH2NLB*. Strikingly, we found that the majority of the population carries three or four *NOTCH2NLA*-derived Exon1$^{A\text{-}(Low)}$-variants, which is associated with a substantial reduction in NOTCH2NL protein level. The fact that about 40% of individuals lack the *NOTCH2NLB*-derived Exon1$^{B\text{-}(High)}$-variant completely could indicate that the high level of NOTCH2NL protein producing variant is slowly being purged from the human genome. We found that this is not the only evolutionary force at play: Next to the Exon1 variants, there are two other deleterious variants, R113* and Exon$^{2B\text{-}(Splice\text{-}mut)}$, that reduce the dosage of functional NOTCH2NL protein. Remarkably, these deleterious variants are more often found in individuals with higher Exon1$^{B\text{-}(High)}$ frequency, indicating that they provide complementary genetic strategies to decrease NOTCH2NL dosage. The R113* and Exon$^{2B\text{-}(Splice\text{-}mut)}$ variants are exclusively present in modern humans and are therefore human-specific adaptations that result in reduced NOTCH2NL protein levels. The driving force behind the evolutionary trend to lower levels of NOTCH2NL protein remains elusive. Phylogenetic comparisons or dN/dS analysis are traditionally used to assess if such variation is significantly associated with evolutionary selection. Because of the absence of functional nonhuman orthologs required to do these comparisons, it is not possible to apply these approaches for analysis of *NOTCH2NL* genes. In addition, frequent and ongoing gene conversion between *NOTCH2* and *NOTCH2NL*-containing loci also hampers this analysis when trying to make comparisons with the truncated *NOTCH2NL* pseudogenes in chimpanzee and gorilla. The high frequency of multiple variants that decrease the available levels of NOTCH2NL protein suggests that *NOTCH2NL* genes have been under selection to counteract high levels of NOTCH2NL expression. Whereas a high frequency of loss-of-function alleles in a population could in principle argue against an essential function of the gene in question and could progress to a complete loss of functional alleles in the future, our data indicate that this is not the case for *NOTCH2NL* genes: Based on the high frequency of loss-of-function variants in *NOTCH2NL* genes in modern humans, it would be expected that a decent proportion of the population would have a genomic configuration without any functional *NOTCH2NL* allele. This is clearly not the case, as the skewed allele distributions that we report points toward purifying selection in order to maintain at least one functional copy of *NOTCH2NL*. This suggests that in present day humans, a certain minimal level of NOTCH2NL protein is required for normal human development. The observed evolutionary changes in *NOTCH2NL* composition could be the result of evolutionary adaptations that took place in any of the tissues where *NOTCH2NL* is expressed, including the developing brain. Even though this remains speculative at the moment, the trend toward lower levels of NOTCH2NL proteins in the human lineage could be correlated to previous observations suggesting a progressive reduction of human brain size that started about 60,000 years ago (Henneberg 1988; Bednarik 2014).

Effectively, NOTCH2NL dosage, which is the total of protein produced by all *NOTCH2NL* loci, may vary between individuals but seems to stay within certain upper and lower ranges. Our new insights regarding the effect of Exon1 variants on NOTCH2NL protein levels may also help in understanding to what extent *NOTCH2NL* genes contribute to *1q21.1* Copy-number variation (CNV)-related phenotypes. Specifically for NOTCH2NL-mediated effects, like potentiating NOTCH signaling, CNVs of an allele carrying the Exon1$^{B\text{-}(High)}$ variant may have a much larger effect than CNVs of an allele carrying the Exon1$^{A\text{-}(Low)}$ variant. Identifying which *NOTCH2NL* loci are affected by gain and loss of alleles will have to be complemented by distribution analysis of Exon1$^{A\text{-}(Low)}$, Exon1$^{C\text{-}(X\text{-}low)}$, R113*, and Exon$^{2B\text{-}(Splice\text{-}mut)}$ variants as they are major determinants of NOTCH2NL levels. The realization that gene conversion between functionally different *NOTCH2NL* genes can contribute to the rapid adaptation of the human species to establish lower levels of NOTCH2NL protein, may prove to be an example for other unstable loci that are characterized by recent segmental duplications. As some of these, like the *1q21.1* locus, are associated with disease, it will be intriguing to see if gene conversion also affects genetic configurations of such loci.

Ever since the availability of genomes from Neanderthals, Denisovans, and ancient humans, the question was raised which modern human-specific coding variants may hold clues to how our species evolved over the last ~600,000 years. Here, we discovered such variants in the *NOTCH2NL* genes, a gene family that emerged in humans about 4 Ma. The role of *NOTCH2NL* genes in human brain development and their involvement in *1q21.1* CNVs associated with a wide variety of neurological disorders emphasizes the importance of the discoveries we describe here: Even if the driving forces of the observed evolutionary changes lie outside the brain, the recent and ongoing structural evolution of human *NOTCH2NL* genes suggests that the tightly coordinated process of human cortical neurogenesis is still subject to fine-tuning.

Materials and Methods

NOTCH2NL Copy Number Analysis from Whole-Genome Sequencing Data

Fastq files were imported from the EBI SRA to the Galaxy EU or US server (Afgan et al. 2018). For Simons diversity data, only the R1 data were used. Reads were trimmed using Trimmomatic (Galaxy v0.36.5) by the following settings: SLIDINGWINDOW: 4, 20 and MINLEN: 30. The remaining reads were mapped to the *NOTCH2NL*-masked hg38 reference genome using Bowtie2 (Galaxy v2.3.4.2), using single-end, very sensitive end-to-end settings. Sequence read depth per genome was ~15–30×. The BAM output files were sliced using samtools slice (Galaxy v2.0.1) with the coordinates chr1:118911553–121069626. Bedtools coverage (Galaxy v2.27.0.2) was applied to each sliced BAM file, reporting coverage for each position. The *NOTCH2*-single-copy region used is located at chr1:119908310–119989035, the *NOTCH2* + *NOTCH2NL* multicopy region used is located at chr1:119990490–120087745. Each region was filtered for repeats using RepeatMasker, and only the nonrepeat intervals were used in coverage analysis. Mean coverage across both regions was calculated by averaging coverage per position. The mean coverage of the *NOTCH2* + *NOTCH2NL*-multicopy region was divided by the mean coverage of the *NOTCH2*-single-copy region to infer *NOTCH2NL* copy-number pet data set. BAM file data were visualized in the UCSC genome browser (Kent et al. 2002). For ancient DNA data sets which consisted of multiple libraries, each library was mapped separately and then merged. The Denisova3 run ERR141700 was omitted due to high sequence duplication. The following WGS data sets were used:

For comparisons of the SUN analysis with previously assembled *NOTCH2NL* configurations (Fiddes et al. 2018), the following samples and data sets were used (Steinberg et al. 2014; Zook et al. 2016; Eberle et al. 2017; Regier et al. 2018; Audano et al. 2019; Marks et al. 2019):

NA24143: 10× genomics (GIAB), WGS (PRJNA200694), WXS (PRJNA200694)

NA24149: 10× genomics (GIAB), WGS (PRJNA200694), WXS (PRJNA200694)

NA24385: 10× genomics (GIAB), WGS (PRJNA200694, PRJNA428496), WXS (PRJNA200694)

NA19240: WGS (PRJNA288807, PRJNA428496, PRJEB4252)

NA12891: WGS and 10× WGS (PRJEB3381, PRJNA428496, PRJNA393319)

CHM1: WGS (PRJNA246220, PRJNA176729)

Separation of *NOTCH2NL* Copy Number per Allele Using SUNs

Based on the hg38 reference genome, single-nucleotide variants and indels were identified, via DNA sequence alignment of the *NOTCH2NLA*, *-B*, *-C*, or *-R* loci to the *NOTCH2* locus. Only SUNs within the region chr1:119990474–12008798 were considered, as this is the maximal duplicon size present in each of the *NOTCH2NL* loci based on the segmental duplication track in the UCSC genome browser hg38. The position of each of these SUNs per locus was stored in BED format. These

Modern human	
PRJEB9586 (Mallick et al. 2016)	Simons diversity genomes
NA (Van Hout et al. 2019)	UK Biobank exomes
Ancient human	
PRJEB6622 (Fu et al. 2014)	Ust'-Ishim
PRJEB6272 (Lazaridis et al. 2014)	Loschbour, StuttgartLBK, Motala3, Motala12
PRJNA240906 (Gamba et al. 2014)	NE1, BR2, IR1, KO1, NE6, NE7, CO1, NE5, BR1
PRJEB4604 (Schiffels et al. 2016)	12880A, 12881A, 12883A, 12884A, 15594A-sc-20
PRJEB21878 (Skoglund et al. 2017)	I9028, I9133, I9134
PRJEB11004 (Martiniano et al. 2016)	3DRIF-16, 3DRIF-26, 6DRIF-18, 6DRIF-21, 6DRIF-22, 6DRIF-23, 6DRIF-3, M1489, NO3423
PRJEB24629 (de la Fuente et al. 2018)	IPK12, IPY10
PRJEB27628 (Krzewińska et al. 2018)	chy002, kzb002, kzb005, kzb006, kzb007, kzb008, mur003, mur004, scy009, scy301, scy303
PRJEB13123 (Fu et al. 2016)	Karelia
PRJEB11364 (Jones et al. 2015)	Bichon, Kotias, Satsurblia
PRJEB21940 (Günther et al. 2018)	Sf12, H22, Sf913, Stg001
PRJEB9783 (Günther et al. 2015)	atp002, atp12-1240
PRJNA218466 (Raghavan et al. 2014)	Mal'Ta
PRJEB21037 (Saag et al. 2017)	Kunila1, Ardu2
PRJEB18067 (Jones et al. 2017)	Latvia_HG1, Latvia_HG2, Latvia_HG3, Latvia_MN2
PRJEB11995 (Cassidy et al. 2016)	BA64, RM127, RSK1, RSK2
PRJEB29663 (Wright et al. 2018)	MH8
PRJEB31045 (Sánchez-Quinto et al. 2019)	ans017, prs016, prs002, prs009
PRJNA338374 (Bhattacharya et al. 2018)	Atacama
PRJEB23467 (Valdiosera et al. 2018)	atp002, atp016
PRJEB7618 (Seguin-Orlando et al. 2014)	Kostenki 14
PRJNA284124 (Rasmussen et al. 2015)	Kennewick
PRJNA46213 (Rasmussen et al. 2010)	Saqqaq
PRJNA229448 (Rasmussen et al. 2014)	Anzick-1
PRJEB6943	Cr10-sc, PA38-sc, PA30-sc
PRJEB2830 (Keller et al. 2012)	Ötzi
PRJNA230689 (Olalde et al. 2014)	La Brana
PRJEB6090 (Skoglund et al. 2014)	Gökhem2, Ajvide58
Neanderthal	
PRJEB1265 (Slon et al. 2017)	Altai
PRJEB21157 (Prüfer et al. 2017)	Vindija
PRJEB21195 (Prüfer et al. 2017)	Mezmaiskaya1
NA (Mafessoni et al. 2020)	Chagyrskaya
PRJEB21870 (Hajdinjak et al. 2018)	Goyet Q56-1
PRJEB21875 (Hajdinjak et al. 2018)	Les Cottes Z4-1514
PRJEB21881 (Hajdinjak et al. 2018)	Mezmaiskaya2
PRJEB21882 (Hajdinjak et al. 2018)	Vindija 87
PRJEB21883 (Hajdinjak et al. 2018)	Spy 94a
PRJEB2065 (Green et al. 2010)	Vi33.16, Vi33.25, Vi33.26
Denisova	
PRJEB3092 (Meyer et al. 2012)	Denisova3
PRJEB20653 (Slon et al. 2017)	Denisova2
Neanderthal/Denisova hybrid	
PRJEB24663 (Slon et al. 2018)	Denisova11

(continued)

were used to generate .vcf format data per BAM file reporting the total read depth and variant (SUN) depth for these positions. This was done using samtools (v1.7) mpileup:

```
samtools mpileup -uvf hg38.fasta -t
DP -t AD -l variant_positions.bed -Q
13 -q 0 -b datasets.txt > output.vcf
```

The relevant information to calculate SUN frequency per allele was extracted using bcftools (v1.7) query:

```
bcftools query -f '%CHROM\t%POS\t%
REF\t%ALT{0} [\t%DP\t%AD{0}\t%AD
{1}]\n' -H mpileup_output.vcf >
mpileup_output_variants.vcf
```

The frequency per variant was calculated using these output files by dividing allele depth for each SUN (AD) by total depth (DP). For each locus, only SUNs with >0.67 frequency in the population were used for analysis to account for ambiguous or population-specific sites that may skew allele distribution calculation, such as known common SNPs. The frequency of the selected SUNs was averaged per locus and multiplied by the total number of alleles calculated previously based on sequence read coverage, to transform allele frequencies into allele counts. Since there are many SUNs for *NOTCH2*, *NOTCH2NLR*, and *NOTCH2NLC*, they provide an accurate estimation for the allele count of these loci. For *NOTCH2NLA* and *NOTCH2NLB*, only a few SUNs are present and gene conversion phenomena happen frequently, which makes this procedure challenging. Therefore, to analyze these loci, we first subtracted the *NOTCH2*, *NOTCH2NLR*, and *NOTCH2NLC* allele counts from the total allele count. The remaining alleles must be derived from *NOTCH2NLA* and *NOTCH2NLB*, and so, the remaining alleles were counted using the ratio of the average SUN frequency for *NOTCH2NLA* and *NOTCH2NLB*. These data were plotted in donut-charts using LibreOffice v6.1.0.3. For graphs showing the per-SUN allele count across the *NOTCH2NL* loci, the *NOTCH2NLB* SUN count was inferred from the *NOTCH2NLA* SUN count in the 5′ region of the locus, where no *NOTCH2NLB* SUNs are present. For example in modern humans there are four *NOTCH2NLA*+*NOTCH2NLB* loci, then the Exon1$^{B\text{-}(High)}$ allele count was calculated according to this: Exon1$^{B\text{-}(High)}$ allele count = 4 − Exon1$^{A\text{-}(Low)}$ count. Correction for NOTCH2 > NOTCH2NLR gene conversion was done for genomes that showed three alleles NOTCH2. These showed a concordant decrease of one allele NOTCH2NLR based on both the coverage analysis and SUN analysis. This difference was corrected for, in example, three alleles NOTCH2 and one allele NOTCH2NLR in one individual were corrected to two alleles NOTCH2 and two alleles NOTCH2NLR. For separation of the Simons diversity genomes data per population, the sample metadata supplied with the data were used.

Position	Locus	Orientation	Reference sequence hg38
chr1:120069403–120069404	NOTCH2	−	ATG
chr1:120724179–120724180	NOTCH2NLR	+	ATG
chr1:146228778–146228779	NOTCH2NLA	−	ATA
chr1:148679531–148679532	NOTCH2NLB	−	ATG
chr1:149390853–149390854	NOTCH2NLC	+	ATG

Position	Locus	Orientation	Reference sequence hg38
chr1:120029988–120029989	NOTCH2	−	T
chr1:120763625–120763626	NOTCH2NLR	+	A
chr1:146189382–146189383	NOTCH2NLA	−	T
chr1:148640098–148640099	NOTCH2NLB	−	C
chr1:149430931–149430932	NOTCH2NLC	+	A

Position	Locus	Orientation	Reference sequence hg38
chr1:119997052–119997053	NOTCH2	−	T
chr1:120793439–120793440	NOTCH2NLR	+	A
chr1:146156535–146156536	NOTCH2NLA	−	T
chr1:148607465–148607466	NOTCH2NLB	−	T
chr1:149463769–149463770	NOTCH2NLC	+	A

Allele Frequencies in UK Biobank Exome Data

Reads mapping on *NOTCH* or *NOTCH2NL* genes were extracted from UK Biobank CRAM exome files (>20× coverage) mapped on hg38. As in these data sets, the reads are mapped to *NOTCH* and all *NOTCH2NL* loci in hg38, the analysis was adjusted from the original analysis that used the masked hg38. For the Exon1$^{A\text{-}(Low)}$ variant (AT**G**>AT**A**), the following positions were analyzed:

Similarly, the Exon$^{2B\text{-}(Splice\text{-}mut)}$ variant information was derived from the following positions:

Nea1^{N232S} variant (AAT > AGT) information was derived from the following positions:

Read depth and allele depth analysis using samtools and bcftools was then done for each locus with the following parameters:

```
samtools mpileup -uvf hg38.fasta -t
DP -t AD -l variant_positions.bed -Q
13 -q 0 -b datasets.txt > output.vcf

bcftools query -f '%CHROM\t%POS\t%
REF\t%ALT{0} [\t%DP\t%AD{0}\t%AD
```

```
{1}]\n' -H output.vcf > query_
output.vcf
```

The setting -q (mapping quality) was set to 0, to include multimapping reads that cannot be confidently assigned to a specific *NOTCH2NL* locus but still contain information regarding variant frequencies. Since the Exon1$^{A\text{-}(Low)}$ variant is annotated in the hg38 genome in *NOTCH2NLA*, reads containing this variant will map there with a better alignment score. As such, the Exon1$^{A\text{-}(Low)}$ frequency was calculated by read depth at the *NOTCH2NLA* position divided by the sum of read depths at the *NOTCH2* + all *NOTCH2NL* loci. The Exon$^{2B\text{-}(Splice\text{-}mut)}$ frequency was calculated by read depth at the specific *NOTCH2NLB* position, where this variant is annotated in hg38, divided by the total read depth at the paralogous positions.

Cell Culture

HEK293 cells (ATCC CRL-1573) were cultured in Dulbecco's modified eagle medium (DMEM) + GlutaMax, high glucose (Thermofisher 61965026), supplemented with 10% heat-inactivated fetal bovine serum (HIFBS) (Thermofisher 10500064) and 100 μg/ml Pen/Strep (Thermofisher 15140122). U2OS and U2OS-JAG2 cells (gifts of Arjan Groot and Marc Vooijs, MAASTRO Lab, Maastricht University) were cultured in DMEM + GlutaMax, high glucose, supplemented with 10% HIFBS and 100 μg/ml Pen/Strep. U2OS-JAG2 cells were additionally supplemented with 2 μg/ml puromycin (Sigma P8833). For routine passaging, medium was removed and cells washed once with phosphate-buffered saline (PBS) (Thermofisher 10010056). A sterile filtered 0.25% Trypsin (Thermofisher 15090046) + 0.5 mM disodium-ethylenediaminetetraacetic acid (EDTA) (Sigma E5134) solution in PBS was added, and incubated at 37 °C for 2 min. One-eighth of the cell suspension was transferred to a new culture vessel of the same size.

Transfection for NOTCH2NL Variant Protein Analysis

HEK293 cells were seeded 24 h before transfection in a six-well plate. One hour before transfection, medium was replaced with 1,800 μl DMEM + GlutaMAX, high glucose and 10% HIFBS. The transfection mix per well was as follows: 500 ng of pCAGN1-NOTCH2NL or pCAGN1-EV, and 500 ng of pCAGEN-GFP were mixed in a total volume of 100 μl 0.25 M $CaCl_2$, after followed by addition of 100 μl 2× HEPES-buffered saline (50 m HEPES, 1.5 mM Na_2HPO_4, 140 mM NaCl, pH 7.05). The 200 μl solutions were mixed by pipetting up and down five times, and the complete mix was added to one well of a six-well plate. Six hours after adding transfection mixes, medium was replaced.

Protein Isolation

Cells were isolated for protein extraction 24–30 h after transfection. Cells were washed twice in ice-cold PBS, then detached using a cell scraper (VWR 734-1527) and transferred to 1.5-ml microcentrifuge tubes. Cell suspensions were centrifuged at 4 °C for 5 min at 1,000 rcf to pellet cells. The supernatant was removed and the cells resuspended in 10× the pellet volume (100–150 μl) of immunoprecipitation lysis buffer (50 mM Tris–HCl pH8.0, 150 mM NaCl, 5 mM $MgCl_2$, 0.5 mM EDTA, 0.2% NP40 substitute, and 5% glycerol), supplemented with 1× protease inhibitor cocktail (Sigma 5892791001). After incubating for 1 h at 4 °C, cell suspensions were transferred through a 273/4 gauge needle ten times and centrifuged at 20,817 rcf for 10 min at 4 °C to pellet cell debris. The supernatant was transferred to a new 1.5-ml microcentrifuge tube and stored at −80 °C.

Protein Gel Electrophoresis and Western Blot

Twenty microliters of protein extract was mixed with 20 μl of 2× laemmli sample buffer (Biorad 1610737) + 50 mM DTT (Sigma D0632). Samples were heated for 5 min at 95 °C and briefly centrifuged. Twenty microliters per sample was loaded on a 1.5-mm poly-acrylamide gel, consisting of two parts. The running gel (12% acrylamide/Bis, 375 mM Tris–HCl pH 8.8, 0.1% ammonium persulfate [APS], 0.1% sodium dodecyl sulfate [SDS], and 0.04% tetramethylethylenediamine [TEMED]) and the stacking gel (5% acrylamide/Bis, 0.125 mM Tris–HCl pH 6.8, 0.1% APS, 0.1% SDS, and 0.1% TEMED). Twenty microliters of sample was loaded per well and 5 μl of marker (Thermofisher #26619) was used for reference. Electrophoresis was done in 25 mM Tris + 192 mM glycine buffer (Biorad 1610771) and 0.1% SDS. Protein was transferred to nitrocellulose membrane (Sigma 10600004), at 100 V for 2 h in Towbin buffer (25 mM Tris, 192 mM glycine, and 20% methanol). Blots were rinsed three times with demi-water, and transfer was checked by ponceau S staining. Blots were rinsed once in Tris buffered saline (20 mM Tris, pH 7.5, 150 mM NaCl) + 0.1% Tween (TBS-T), followed by incubation in blocking buffer (TBS-T + 5% w/v skim milk powder) for 90 min at room temperature on a shaking platform. Primary antibodies were incubated overnight at 4 °C in TBS-T in 50-ml tubes on a tube roller. Antibodies used were rabbit anti-HA tag (1:6,000, Abcam ab9110) or rabbit anti-GFP (1:4,000, Abcam ab290). Blots were rinsed once in TBS-T and washed in TBS-T three times 15 min on a shaking platform. Secondary antibody goat anti-rabbit-HRP in TBS-T (1:20,000, Thermofisher 656120) was incubated for 60 min at room temperature. Blots were rinsed once in TBS-T and washed three times 15 min in TBS-T on a shaking platform. The SuperSignal Westdura substrate (Thermofisher 34075) was used for chemiluminescent detection, imaged with a ChemiDoc MP imaging system (Biorad 1708280). Signals were quantified using Fiji ImageJ using the NOTCH2NL/GFP ratio.

Coculture NOTCH Reporter Assay

To monitor modulation of NOTCH2 activity by NOTCH2NL, a reporter assay was used. The pGL3-UAS luciferase reporter can be activated by S3-cleaved NOTCH2-Gal4-N1TAD receptor intracellular domain (Gal4 domain fused to NOTCH1-transactivation domain) (gifts of Arjan Groot and Marc Vooijs, MAASTRO Lab, Maastricht University). To achieve high levels of receptor activation, the cells transfected with pcDNA5-NOTCH2-Gal4-N1TAD are cocultured with JAG2 expressing cells. Coculture with regular U2OS cells was

done as a control. pCAGN1-EV or pCAGN1-NOTCH2NL (derived from Addgene 51142) were cotransfected to measure effects of NOTCH2NL on reporter activity. pRL-CMV (Promega E2261) was used for normalization.

For transfection, U2OS cells were seeded in six-well plates at a density of 400,000 cells per well. For coculture assay, U2OS cells or U2OS-JAG2 cells were seeded in 12-well plates at a density of 110,000 cells per well. Twenty-four hours later, U2OS cells in six-well plates were transfected. The transfection complex per well was made by adding 2,500 ng plasmid DNA mix, as described in the table below, in 100 μl OptiMEM (Thermofisher 31985047). In a different tube, 8.33 μl PEI (1 mg/ml, Polysciences 23966) was added to 100 μl OptiMEM. One hundred microliters of each mix was combined, incubated 20 min at room temperature, and added to the well containing 2 ml of complete medium. Reactions were scaled accordingly to facilitate large-scale transfections. Six hours after transfection, the transfected cells were replated onto the 12-wells plate for coculture with U2OS or U2OS-JAG2 cells. Per well, medium was removed and cells were washed once with 1 ml PBS. Trypsin-EDTA (0.5 ml) in PBS was added plates were incubated 90 s at 37 °C. Two millilietrs of complete medium was added, and cell aggregates were broken up by pipetting up and down three times. Cell suspension was transferred to 15-ml conical tubes already containing 4.5 ml of complete medium. From the 12-well plates, the medium was removed and replaced by 1 ml of cell suspension. To control wells, 1 μl of 200 μM DBZ was added. Twenty-four hours after replating, the cells were isolated for luciferase assays using Dual-Luciferase Reporter Assay System (Promega E1980). Medium was removed and each well washed once with 0.5 ml PBS. A total of 150 μl of 1× passive lysis buffer (Promega E1941) was added per well and incubated 15 min on a rotating platform. Plates were wrapped in parafilm and stored at −80 °C. For analysis, 20 μl sample was pipetted to a 96-well optiplate (PerkinElmer 6005290). Samples were measured on a GloMax Navigator device (Promega GM2010), with the following settings: Injector 1, LARII buffer (volume 50 μl, speed 200 μl/s). Wait 2 s. Measure luminescence Luciferase (integration 10 s, readings 1, interval 0.3 s). Injector 2, Stop & Glo buffer (volume 50 μl, speed 200 μl/s). Wait 2 s. Measure luminescence Renilla (integration 10 s, readings 1, interval 0.3 s). For comparison of human, Neanderthal and Denisovan NOTCH2NL variants, the 48 ng pCAGN1-NOTCH2NL condition was used.

	6	16	48	140	420
pGL3-UAS	1,050	1,050	1,050	1,050	1,050
pRL-CMV	70	70	70	70	70
pCAGEN-GFP	35	35	35	35	35
pcDNA5-NOTCH2-Gal4-N1TAD	21	21	21	21	21
pCAGN1-EV/-NOTCH2NL	5/6	14/16	41/48	120/140	361/420
pBluescript (EV/NOTCH2NL)	1,315/1,314	1,306/1,304	1,279/1,272	1,200/1,180	959/900

Plasmids
pCAGEN-GFP (Addgene #11150)
pCAGN1- hCas9 (Addgene #51142)
pCAGN1- EV
pCAGN1-NOTCH2NL
pCAGN1-NOTCH2NL-T197I
pCAGN1-NOTCH2NL-M40I, T197I
pCAGN1-NOTCH2NL-N232S, T197I
pCAGN1-NOTCH2NL-E258A, T197I
pCAGN1-NOTCH2NL-M1I
pCAGN1-NOTCH2NL-M1I, T197I
pCAGN1-NOTCH2NL-M1I, M40I
pCAGN1-NOTCH2NL-M1I, N232S
pCAGN1-NOTCH2NL-M1I, E258A
pCAGN1-NOTCH2NL-HA
pCAGN1-NOTCH2NL-HA-T197I
pCAGN1-NOTCH2NL-HA-M40I, T197I
pCAGN1-NOTCH2NL-HA-N232S, T197I
pCAGN1-NOTCH2NL-HA-E258A, T197I
pCAGN1-NOTCH2NL-HA-M1I
pCAGN1-NOTCH2NL-HA-M1I, T197I
pCAGN1-NOTCH2NL-HA-M1I, M40I
pCAGN1-NOTCH2NL-HA-M1I, N232S
pCAGN1-NOTCH2NL-HA-M1I, E258A
pCAGN1-NOTCH2NL-HA-5′ M1
pCAGN1-NOTCH2NL-HA-5′ M1 + kozak
pCAGN1-NOTCH2NL-HA-5′ M40
pCAGN1-NOTCH2NL-HA-5′ M40 + kozak
pCAGN1-NOTCH2NL-HA-5′ M84
pCAGN1-NOTCH2NL-HA-5′ M84 + kozak
pCAGN1-NOTCH2NL-HA-5′ M1I-I1
pCAGN1-NOTCH2NL-HA-5′ P2
pCAGN1-NOTCH2NL-HA-5′ L12
pCAGN1-NOTCH2NL-HA-5′ P22
pCAGN1-NOTCH2NL-HA-5′ C28
pCAGN1-NOTCH2NL-HA-M1I-ΔI1
pCAGN1-NOTCH2NL-HA-M1I-ΔL4
pCAGN1-NOTCH2NL-HA, 5′ M1, $CTG^{(1-5)} > CTA^{(1-5)}$
pCAGN1-NOTCH2NL-M1I-HA, 5′ I1, $CTG^{(1-5)} > CTA^{(1-5)}$
pCAGN1-NOTCH2NL-M1I-HA, 5′ I1, Δata-$CTG^{(1-5)} > CTA^{(1-5)}$
pcDNA5-NOTCH2-GAL4-TAD-N1
pRL-CMV (Promega E2261)

Amount of plasmid DNA (ng) transfected per condition. pCAGN1-EV/pCAGN1-NOTCH2NL denotes amount of plasmid used per condition accounting for molarity. pBluescript amount was adjusted accordingly as well.

In Silico Analysis of Archaic Coding Variants

For multiple sequence alignment of NOTCH1, -2, and -3 EGF-L domains 6, the relevant sequences were acquired from UniProt and compared using the alignment tool of UniProt. The EGF-L repeat domain consensus sequence was retrieved from Prosite: PDOC00021, EGF_3 PS50026. For MutPred2 and IUPred2A analysis, the archaic amino acid variants were introduced in the NOTCH2NLB protein sequence retrieved from UniProt (P0DPK3). MutPred2 was run with a *P* value threshold of 0.05. IUPred2A was used

with the following settings: Long disorder, Context-dependent predictions: ANCHOR2.

Statistics

Luciferase reporter assay data were first analyzed using analysis of variance (ANOVA) by the R function "aov()." Significant groups were further tested with Tukey's test using the R function "TukeyHSD()." Western blot data were analyzed in the same way, except for data presented in figure 3*B* and *C*, which showed unequal variance (Levene test $P = 0.002$) and were analyzed instead using Welch corrected ANOVA using the R functions "oneway()" with parameters "levene=TRUE" and "corrections=TRUE," followed by Games–Howell test from function "posthocTGH()," with parameter "method=games-howell" (R package "userfriendlyscience"). Population genetic data from Simons diversity genomes and UK Biobank exomes were first analyzed using Kruskal–Wallis tests by the R function "kruskal.test()."Significant groups were further tested with Dunn's test, using the "dunn.test()" function (R package "dunn.test"). Distributions in figure 5C and supplementary figure S5*A*, Supplementary Material online, were tested using the Kolmogorov–Smirnov test using the "ks.test()" function. Expected distributions were generated using the "rnorm()" function in R. For Simons data, this was simulated by generating mean allele counts according to an AABB × AABB polygenic inheritance pattern for the Exon1A-(Low) and Exon1B-(High) variants: 0 alleles (1/16), 1 allele (4/16), 2 alleles (6/16), 3 alleles (4/16), or 4 alleles (1/16), total $N = 2{,}790$. Standard deviation was set to 0.34 to introduce sampling variation. Expected distribution in the UK Biobank analysis was done similarly, except using allele frequencies instead, of 0, 0.1, 0.2, 0.3, or 0.4, total $N = 50{,}000$, with a standard deviation of 0.034, which were adjusted for loss of *NOTCH2NLC* and *NOTCH2NLR* as identified in Simons diversity genomes. Boxplots show median and interquartile range (25th and 75th percentiles), whiskers are defined by 1.5 × interquartile range. Outliers were hidden in violin/box plots from figure 6 and supplementary figures S6–S7, Supplementary Material online, to avoid clutter. All *P* values shown were adjusted for multiple testing using Holm's method.

Data Visualization

For donut-charts showing *NOTCH2NL* allele counts, LibreOffice v6.1.0.3 was used. Data involving genomic context were visualized on the UCSC genome browser and exported as.pdf files. Plots showing quantification of sequence read coverage, luciferase assays, Western blots, per-SUN count graphs, variant allele counts, and distributions were generated in RStudio v1.1.463 and R v3.5.3 with the ggplot2 package v3.1.0. Fig. panels were assembled in Adobe Illustrator v23.0.3.

Data Availability

All genomics data from the Simons Diversity Cohort and Archaic genomes were downloaded from their original depositories. Accession numbers and unique identifiers are provided where necessary. All data from the analyses in this manuscript are included in this published article (and its Supplementary Material online). The raw data from the UK Biobank are not publicly available due to restrictions, but the analyzed data as described in this manuscript are available upon request from the corresponding author. Please note that for UKB analyses we can only share summarized data. Individual-level data may be accessed by submitting an application to UKB.

Code Availability

All code and software used in this manuscript are described and/or available in the Materials and Methods section.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Danielle Posthuma (VU, Amsterdam); Marco Hoekman and Marten Smidt (SILS, University of Amsterdam) for helpful discussions; the Simons Diversity Project and UK Biobank for human genome data; Arjan Groot and Marc Vooijs (Maastricht Radiation Oncology Lab) for reagents for Notch reporter assays; the Galaxy Project for use of their servers to reanalyze part of the WGS data; and people from the Jacobs lab for helpful discussion and comments on the manuscript. This research has been conducted using the UK Biobank resource under application number 16406. This work was supported by ERC starting grant ERC-2016-StG-716035 (F.M.J.J.).

Author Contributions

F.M.J.J. and G.A.L. conceptualized the study and performed the methodology; G.A.L., I.V., and J.E.S. validated the study; F.M.J.J. and G.A.L. performed the investigation; F.M.J.J., G.A.L., J.E.S., and D.P.F. performed the data curation; F.M.J.J. and G.A.L. wrote the original draft; F.M.J.J., G.A.L., and D.P.F. reviewed and edited the manuscript; G.A.L. and D.P.F. performed the visualization; and F.M.J.J. performed the supervision, project administration, and funding acquisition.

References

Afgan E, Baker D, Batut B, Van Den Beek M, Bouvier D, Ech M, Chilton J, Clements D, Coraor N, Grüning BA, et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46(W1):W537–W544.

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AME, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the major structural variant alleles of the human genome. *Cell* 176(3):663–675.e19.

Bednarik RG. 2014. Doing with less: hominin brain atrophy. *HOMO* 65(6):433–449.

Bernier R, Steinman KJ, Reilly B, Wallace AS, Sherr EH, Pojman N, Mefford HC, Gerdts J, Earl R, Hanson E, et al. 2016. Clinical phenotype of the recurrent 1q21.1 copy-number variant. *Genet Med.* 18(4):341–349.

Bhattacharya S, Li J, Sockell A, Kan MJ, Bava FA, Chen SC, Avila-Arcos MC, Ji X, Smith E, Asadi NB, et al. 2018. Whole-genome sequencing of Atacama skeleton shows novel mutations linked with dysplasia. *Genome Res.* 28(4):423–431.

Bruder CEG, Piotrowski A, Gijsbers A, Andersson R, Erickson S, Diaz de Ståhl T, Menzel U, Sandgren J, von Tell D, Poplawski A, et al. 2008.

Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am J Hum Genet*. 82(3):763–771.
Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, Sahoo T, Lalani SR, Graham B, Lee B, Shinawi M, et al. 2008. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet*. 40(12):1466–1471.
Cassidy LM, Martiniano R, Murphy EM, Teasdale MD, Mallory J, Hartwell B, Bradley DG. 2016. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc Natl Acad Sci U S A*. 113(2):368–373.
Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*. 8(10):762–775.
Dannemann M, Racimo F. 2018. Something old, something borrowed: admixture and adaptation in human evolution. *Curr Opin Genet Dev*. 53:1–8.
de la Fuente C, Ávila-Arcos MC, Galimany J, Carpenter ML, Homburger JR, Blanco A, Contreras P, Cruz Dávalos D, Reyes O, San Roman M, et al. 2018. Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia. *Proc Natl Acad Sci U S A*. 115(17):E4006–E4012.
Deng J, Gu M, Miao Y, Yao S, Zhu M, Fang P, Yu X, Li P, Su Y, Huang J, et al. 2019. Long-read sequencing identified repeat expansions in the 5′UTR of the NOTCH2NLC gene from Chinese patients with neuronal intranuclear inclusion disease. *J Med Genet*. 56(11):758–764.
Dougherty ML, Nuttle X, Penn O, Nelson BJ, Huddleston J, Baker C, Harshman L, Duyzend MH, Ventura M, Antonacci F, et al. 2017. The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biol*. 18(1):49.
Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE. 2018. Transcriptional fates of human-specific segmental duplications in brain. *Genome Res*. 28(10):1566–1576.
Duan Z, Li F-Q, Wechsler J, Meade-White K, Williams K, Benson KF, Horwitz M. 2003. A novel notch protein, N2N, targeted by neutrophil elastase and implicated in hereditary neutropenia. *Mol Cell Biol*. 24(1):58–70.
Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, Iqbal Z, Chuang HY, Humphray SJ, Halpern AL, et al. 2017. A reference dataset of 5. 4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res*. 27(1):157–131.
Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, et al. 2018. Human-specific NOTCH2NL genes affect notch signaling and cortical neurogenesis. *Cell* 173(6):1356–1369.e22.
Florio M, Heide M, Pinson A, Brandl H, Albert M, Winkler S, Wimberger P, Huttner WB, Hiller M. 2018. Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *Elife* 7:1–37.
Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, De Filippo C, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514(7523):445–449.
Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A, et al. 2016. The genetic history of Ice Age Europe. *Nature* 534(7606):200–205.
Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, Domboróczki L, Kővári I, Pap I, Anders A, et al. 2014. Genome flux and stasis in a five millennium transect of European prehistory. *Nat Commun*. 5:5257.
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz M-Y, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328(5979):710–722.
Groot AJ, Habets R, Yahyanejad S, Hodin CM, Reiss K, Saftig P, Theys J, Vooijs M. 2014. Regulated proteolysis of NOTCH2 and NOTCH3 receptors by ADAM10 and presenilins. *Mol Cell Biol*. 34(15):2822–2832.
Günther T, Malmström H, Svensson EM, Omrak A, Sánchez-Quinto F, Kılınç GM, Krzewińska M, Eriksson G, Fraser M, Edlund H, et al. 2018. Population genomics of Mesolithic Scandinavia: investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol*. 16(1):e2003703.
Günther T, Valdiosera C, Malmström H, Ureña I, Rodriguez-Varela R, Sverrisdóttir ÓO, Daskalaki EA, Skoglund P, Naidoo T, Svensson EM, et al. 2015. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc Natl Acad Sci U S A*. 112(38):11917–11922.
Habets RAJ, Groot AJ, Yahyanejad S, Tiyanont K, Blacklow SC, Vooijs M. 2015. Human NOTCH2 is resistant to ligand-independent activation by metalloprotease adam17. *J Biol Chem*. 290(23):14705–14716.
Hajdinjak M, Fu Q, Hübner A, Petr M, Mafessoni F, Grote S, Skoglund P, Narasimham V, Rougier H, Crevecoeur I, et al. 2018. Reconstructing the genetic history of late Neanderthals. *Nature* 555(7698):652–656.
Hayashi T, Katagiri S, Mizobuchi K, Yoshitake K, Kameya S, Matsuura T, Iwata T, Nakano T. 2020. Heterozygous GGC repeat expansion of NOTCH2NLC in a patient with neuronal intranuclear inclusion disease and progressive retinal dystrophy. *Ophthalmic Genet*. 41:93–95.
Henneberg M. 1988. Decrease of human skull size in the Holocene. *Hum Biol*. 60:395–405.
Herculano-Houzel S. 2009. The human brain in numbers: a linearly scaled-up primate brain. *Front Hum Neurosci*. 3:31.
Hofman M. 2014. Evolution of the human brain: when bigger is better. *Front Neuroanat*. 8:15.
Holloway RL, Broadfield DC, Yuan MS. 2004. The human fossil record. Hoboken (NJ): John Wiley & Sons, Inc.
Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, Almansour MA, Kikuchi JK, Taira M, Mitsui J, et al. 2019. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat Genet*. 51(8):1222–1232.
Jafar-Nejad H, Leonardi J, Fernandez-Valdivia R. 2010. Role of glycans and glycosyltransferases in the regulation of Notch signaling. *Glycobiology* 20(8):931–949.
Jiao B, Zhou L, Zhou Y, Weng L, Liao X, Tian Y, Guo L, Liu X, Yuan Z, Xiao X, et al. 2020. Identification of expanded repeats in NOTCH2NLC in neurodegenerative dementias. *Neurobiol Aging* 89:142.e1–142.e7.
Johnson MB, Sun X, Kodani A, Borges-Monroy R, Girskis KM, Ryu SC, Wang PP, Patel K, Gonzalez DM, Woo YM, et al. 2018. Aspm knockout ferret reveals an evolutionary mechanism governing cerebral cortical size letter. *Nature* 556(7701):370–375.
Jones ER, Gonzalez-Fortes G, Connell S, Siska V, Eriksson A, Martiniano R, McLaughlin RL, Gallego Llorente M, Cassidy LM, Gamba C, et al. 2015. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat Commun*. 6:8912.
Jones ER, Zarina G, Moiseyev V, Lightfoot E, Nigst PR, Manica A, Pinhasi R, Bradley DG. 2017. The Neolithic transition in the Baltic was not driven by admixture with early European farmers. *Curr Biol*. 27(4):576–582.
Ju XC, Hou QQ, Sheng AL, Wu KY, Zhou Y, Jin Y, Wen T, Yang Z, Wang X, Luo ZG. 2016. The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *Elife* 5:pii: e18197.
Kalebic N, Gilardi C, Albert M, Namba T, Long KR, Kostic M, Langen B, Huttner WB. 2018. Human-specific ARHGAP11B induces hallmarks of neocortical expansion in developing ferret neocortex. *Elife* 7:e41241.
Kearse MG, Wilusz JE. 2017. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev*. 31(17):1717–1731.
Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, et al. 2012. New insights into the Tyrolean Iceman' s origin and phenotype as inferred by whole-genome sequencing. *Nat Commun*. 3:698.
Kent WJ, Sugnet CW, Furey TS, Roskin KM. 2002. The human genome browser at UCSC. *J Med Chem*. 19:1228–1231.

Krzewińska M, Kılınç GM, Juras A, Koptekin D, Chyleński M, Nikitin AG, Shcherbakov N, Shuteleva I, Leonova T, Kraeva L, et al. 2018. Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads. *Sci Adv*. 4(10):eaat4457.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513(7518):409–413.

Liu J, Liu W, Yang L, Wu Q, Zhang H, Fang A, Li L, Xu X, Sun L, Zhang J, et al. 2017. The primate-specific gene TMEM14B marks outer radial glia cells and promotes cortical expansion and folding. *Cell Stem Cell* 21(5):635–649.e8.

Mafessoni F, Grote S, de Filippo C, Slon V, Kolobova KA, Viola B, Markin S V, Chintalapati M, Peyrégne S, Skov L, et al. 2020. A high-coverage Neandertal genome from Chagyrskaya Cave. bioRxiv: 2020.03.12.988956.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538(7624):201–206.

Marino L. 1998. a comparison of encephalization between odontocete cetaceans and anthropoid primates. *Brain Behav Evol*. 51(4):230–238.

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, et al. 2019. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res*. 29(4):635–645.

Martiniano R, Caffell A, Holst M, Hunter-Mann K, Montgomery J, Müldner G, McLaughlin RL, Teasdale MD, van Rheenen W, Veldink JH, et al. 2016. Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nat Commun*. 7:10326.

Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, Huang S, Maloney VK, Crolla JA, Baralle D, et al. 2008. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med*. 359(16):1685–1699.

Mészáros B, Erdős G, Dosztányi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res*. 46(W1):W329–W337.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222–226.

Okubo M, Doi H, Fukai R, Fujita A, Mitsuhashi S, Hashiguchi S, Kishida H, Ueda N, Morihara K, Ogasawara A, et al. 2019. GGC repeat expansion of NOTCH2NLC in adult patients with leukoencephalopathy. *Ann Neurol*. 86(6):962–968.

Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, Prado-Martinez J, Rodríguez JA, Rasmussen S, Quilez J, et al. 2014. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* 507(7491):225–228.

Pejaver V, Mooney SD, Radivojac P. et al. 2017. Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. Hum Mutat. 38:1092–1108.

Pollen AA, Nowakowski TJ, Chen J, Retallack H, Sandoval-Espinosa C, Nicholas CR, Shuga J, Liu SJ, Oldham MC, Diaz A, et al. 2015. Molecular identity of human outer radial glia during cortical development. *Cell* 163(1):55–67.

Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. 2017. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* 358(6363):655–658.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, De Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505(7481):43–49.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Orlando L, Metspalu E, et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505(7481):87–91.

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW Jr, Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, et al. 2014. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* 506(7487):225–229.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463(7282):757–762.

Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, Zollikofer CPE, Ponce de León MS, Allentoft ME, Moltke I, et al. 2015. The ancestry and affiliations of Kennewick Man. *Nature* 523(7561):455–458.

Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, Chen BJ, Kher M, Banks E, Ames DC, et al. 2018. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat Commun*. 9:1–8.

Saag L, Varul L, Scheib CL, Stenderup J, Allentoft ME, Saag L, Pagani L, Reidla M, Tambets K, Metspalu E, et al. 2017. Extensive farming in Estonia started through a sex-biased migration from the steppe. *Curr Biol*. 27(14):2185–2193.e6.

Sánchez-Quinto F, Malmström H, Fraser M, Girdland-Flink L, Svensson EM, Simões LG, George R, Hollfelder N, Burenhult G, Noble G, et al. 2019. Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. *Proc Natl Acad Sci U S A*. 116(19):9469–9474.

Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, Clarke R, Lyons A, Mortimer R, Sayer D, et al. 2016. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat Commun*. 7:10408.

Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V, et al. 2014. Genomic structure in Europeans dating back at least 36,200 years. *Science* 346(6213):1113–1118.

Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, Hall P, Tambets K, Parik J, Sjögren K-G, et al. 2014. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* 344(6185):747–750.

Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, Salie T, Rohland N, Mallick S, Peltzer A, et al. 2017. Reconstructing prehistoric African population structure. *Cell* 171(1):59–71.e21.

Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, Hajdinjak M, Peyrégne S, Nagel S, Brown S, et al. 2018. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* 561(7721):113–116.

Slon V, Viola B, Renaud G, Gansauge M-T, Benazzi S, Sawyer S, Hublin J-J, Shunkov MV, Derevianko AP, Kelso J, et al. 2017. A fourth Denisovan individual. *Sci Adv*. 3(7):e1700186.

Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, Koike H, Hashiguchi A, Takashima H, Sugiyama H, et al. 2019. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat Genet*. 51(8):1215–1221.

Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, Shiryev SA, Morgulis A, Surti U, Warren WC, et al. 2014. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res*. 24(12):2066–2076.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. 2010. Diversity of Human Copy Number Variation and Multicopy Genes. Science 330(6004):641–646.

Sun QY, Xu Q, Tian Y, Hu ZM, Qin LX, Yang JX, Huang W, Xue J, Li JC, Zeng S, et al. 2020. Expansion of GGC repeat in the human-specific NOTCH2NLC gene is associated with essential tremor. *Brain* 143(1):222–233.

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F, et al. 2018. Human-

specific NOTCH2NL genes expand cortical neurogenesis through delta/notch regulation. *Cell* 173(6):1370–1384.e16.
Takeuchi H, Yu H, Hao H, Takeuchi M, Ito A, Li H, Haltiwanger RS. 2017. O-Glycosylation modulates the stability of epidermal growth factor-like repeats and thereby regulates Notch trafficking. *J Biol Chem*. 292(38):15964–15973.
Vadgama N, Pittman A, Simpson M, Nirmalananthan N, Murray R, Yoshikawa T, De Rijk P, Rees E, Kirov G, Hughes D, et al. 2019. De novo single-nucleotide and copy number variation in discordant monozygotic twins reveals disease-related genes. *Eur J Hum Genet*. 27(7):1121–1133.
Valdiosera C, Günther T, Vera-Rodríguez JC, Ureña I, Iriarte E, Rodríguez-Varela R, Simões LG, Martínez-Sánchez RM, Svensson EM, Malmström H, et al. 2018. Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. *Proc Natl Acad Sci U S A*. 115(13):3428–3433.
Van Hout CV, Tachmazidou I, Backman JD, Hoffman JX, Ye B, Pandey AK, Gonzaga-Jauregui C, Khalid S, Liu D, Banerjee N, et al. 2019. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. bioRxiv 572347.
Wright JL, Wasef S, Heupink TH, Westaway MC, Rasmussen S, Pardoe C, Fourmile GG, Young M, Johnson T, Slade J, et al. 2018. Ancient nuclear genomes enable repatriation of Indigenous human remains. *Sci Adv*. 4(12):eaau5064.
Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 3:1–26.